ANÁLISIS CUANTITATIVO
DE DATOS PARA LA INVESTIGACIÓN EDUCATIVA Y SOCIAL

Christian Hederich Martínez

UNIVERSIDAD PEDAGOGICA NACIONAL
Educadora de educadores

Análisis cuantitativo de datos para la investigación educativa y social

Análisis cuantitativo de datos para la investigación educativa y social

Christian Hederich Martínez

UNIVERSIDAD PEDAGOGICA NACIONAL
Educadora de educadores

Catalogación en la fuente–Biblioteca Central de la Universidad Pedagógica Nacional

Hederich Martínez. Christian

Análisis cuantitativo de datos para la investigación educativa Social / Christian Hederich Martínez. – Primera edición. -- Bogotá. Universidad Pedagógica Nacional, 2023.

316 páginas. Representaciones gráficas. (Tablas-cuadros)

Incluye: Referencias bibliográficas
Incluye: Índice analítico

ISBN impreso: 978-628-7518-91-9
ISBN ePub: 978-628-7518-93-3
ISBN PDF: 978-628-7518-92-6

1. Métodos Analíticos. 2. Procesamiento de Datos. 3. Estadística - Programas para Computador. 4. Análisis de Datos – Programas para Computador. 5. Lenguajes de Programación. 6. Paquetes Estadísticos - Educación. 7. Base de Datos - Estadística. I. Tít.

005.36 21. ed.

Análisis cuantitativo de datos para la investigación educativa y social

Autor
Christian Hederich Martínez

ISBN impreso: 978-628-7518-91-9
ISBN ePub: 978-628-7518-93-3
ISBN PDF: 978-628-7518-92-6

Primera edición, 2023

Preparación editorial

Grupo Interno de Trabajo Editorial
Universidad Pedagógica Nacional
Carrera 16A n.º 79-08
editorial.upn.edu.co
Teléfono: (57 1) 347 1190 - (57 1) 594 1894
Bogotá, Colombia

Alba Lucía Bernal Cerquera
Coordinación

Pablo A. Castro Henao
Edición

Karen Grisales
Martha Méndez
Corrección de estilo

Julián Hernández - Taller de diseño
Diagramación

Juan Camilo Corredor
Diseño de tablas y figuras

Paula Andrea Cubillos Gómez
Finalización de artes

Fredy Johan Espitia Ballesteros
Diseño de cubierta

Xpress Estudio Gráfico y Digital, S. A. S./Kimpres
Impresión

Fecha de evaluación: 21-07-2022 / 08-08-2022
Fecha de aprobación: 24-02-2022

Hecho el depósito legal que ordena la Ley 44 de 1993 y el decreto reglamentario 460 de 1995.

UNIVERSIDAD PEDAGOGICA NACIONAL
Educadora de educadores

Contenido

Lista de figuras

Lista de tablas

Agradecimientos

Debo expresar mis agradecimientos a multitud de instituciones y personas que, de forma directa, colaboraron en la elaboración de este libro. En la dimensión de las instituciones, quisiera agradecer a la Universidad Pedagógica Nacional, en cabeza del rector, Dr. Leonardo Fabio Martínez (periodo 2018-2022), y de la decana de la Facultad de Educación, Dra. Sandra Durán, por haberme concedido el privilegio de un año sabático, tiempo en el cual se redactó la mayor parte del presente texto. Tristemente, no fue sino que se me concediera el sabático para que se declararan la pandemia y su respectivo encierro, por lo que no pude avanzar en los viajes que tenía planeados como insumo para la elaboración del escrito; lo que, por otro lado, contribuyó de forma significativa a la longitud del libro y a la multiplicidad de los temas tratados.

Muy especial agradecimiento debo al Dr. Carlos Lanziano Molano, amigo, estadístico y experto en los temas tratados en este texto, quien tiene la rara habilidad de permitir los cuestionamientos, a veces ingenuos, que los neófitos en algunos temas nos atrevemos a hacer, con la apertura mental necesaria para hacerse la misma pregunta y respondérsela con las herramientas que le ha dado su misma experticia. El resultado fue un diálogo fecundo y agradable. El Dr. Lanziano revisó cada uno de los temas y textos contenidos en el libro, y resolvió muchísimas de las preguntas que surgían en el proceso.

También debo agradecer, de manera muy sentida, a los integrantes de nuestro grupo de investigación, el Grupo de Estilos Cognitivos, profesoras Ángela Camargo, Carolina Hernández, Dora Manjarrés y Diana Abello, por su apoyo durante ese año. Por su interés en estos temas, la profesora Abello tuvo la gentileza adicional de leer el manuscrito completo y comentarlo en un acto formal preparado por la Facultad de Educación para presentar el texto.

Por último, debo expresar mi agradecimiento a los profesionales del Grupo Interno de Trabajo Editorial de la Universidad Pedagógica Nacional, y en particular a Lucía Bernal Cerquera, jefe del grupo, y a Pablo Castro Henao, editor de este texto.

Desde el momento en que se contó con una primera versión, el texto fue experimentado en algunas cátedras relacionadas con la estadística y el procesamiento cuantitativo de datos en el Doctorado en Psicología de la Universidad del Norte y en el Doctorado Interinstitucional en Educación de la Universidad Pedagógica Nacional, la Universidad Distrital y la Universidad del Valle. Durante estas sesiones, diferentes alumnos contribuyeron de forma muy importante, haciéndome ver errores y segmentos que resultaban confusos al lector. A estos alumnos les doy las gracias por sus contribuciones, y a todos mis alumnos, los presentes, los pasados y los que aún están por venir, por darle lo principal al texto: un motivo.

Introducción

Propósito y uso de la obra

El presente libro pretende servir de manual y guía en el procesamiento y análisis cuantitativo de datos en procesos de investigación social, con énfasis en investigación educativa. De igual forma, puede usarse como libro de texto en un curso de análisis cuantitativo de datos.

En esos términos, se presenta como un apoyo que se extiende desde el momento en que se confecciona una base de datos para su procesamiento, hasta la presentación final de los resultados en un artículo científico. Así, intenta agrupar, en una única publicación, un conjunto de elementos, trucos, procesos y conocimientos con diversos grados de estructuración, que se requieren, o que facilitan, el procesamiento y análisis cuantitativo de la información.

Es importante notar que hablamos de procesamiento y análisis cuantitativo, y no de procesamiento de datos cuantitativos. La diferencia es evidente. En nuestro caso nos referimos a las técnicas cuantitativas para el análisis de cualquier tipo de datos, sean estos propiamente cuantitativos, y por lo tanto numéricos, o cualitativos —codificados, eso sí, en formatos numéricos—.

Actualmente, para el procesamiento y análisis de datos cuantitativos utilizamos computadores con programas especializados en esas tareas. Pretender hacerlo de otra forma es anacrónico, a menos que el lector esté interesado en los aspectos más detallados de estos procesos. Desde mi experiencia, el tesista en maestría o doctorado no tiene ese interés, pero sí requiere realizar un buen proceso de análisis, que se pueda comprender y explicar correctamente. A este lector va dirigida la obra. Por esta razón, entraremos de lleno en el uso de paquetes estadísticos.

Uso de paquetes estadísticos

Existen muchos programas de computador que resultan útiles para el procesamiento y análisis de datos. Entre los más conocidos están el SPSS, el SAS, Statistica y el Stata, si bien recientemente ha ganado gran popularidad el uso de R, gracias a sus grandes capacidades de graficación y a su uso libre. Más que un programa para el procesamiento, el R es un entorno y un lenguaje de programación orientado al análisis estadístico que forma parte de un proyecto colaborativo y abierto. Estas características le han permitido al R un gran crecimiento y una rápida expansión. Requiere de

cierta dedicación inicial para su dominio, por lo que, aunque recomendamos su uso, no dependeremos de este en el contexto de la presente obra.

Muchos de los procedimientos descritos en esta obra se pueden obtener mediante un adecuado uso de la popular hoja electrónica de Microsoft, MS-Excel. Esta también es una opción válida y confiable para el procesamiento de datos, aunque no siempre resulta ser la más cómoda.

En el presente libro utilizaremos con frecuencia dos paquetes estadísticos. El primero es uno de los paquetes de software más populares para el procesamiento y análisis de datos estadísticos en la información social: el SPSS (Statistical Package for Social Sciences). Este programa, que goza de gran popularidad, es distribuido por IBM (IBM-SPSS) y permite trabajar con grandes bases de datos de una manera sencilla y cómoda. La dificultad para trabajar con este paquete es que no es de uso gratuito. Para esta obra, los ejemplos fueron obtenidos en la versión 27 del programa, corriendo bajo el sistema operativo MacOS Big Sur Versión 14.5.

El IBM-SPSS funciona en tres ventanas separadas: la de base de datos, la de resultados y la del editor de sintaxis. Siempre que esté activa alguna de estas ventanas estará también una barra con menús en la parte superior de la pantalla. La figura 1 muestra la pantalla con una base de datos activa. Las instrucciones para el análisis se especifican en la barra superior. Los resultados del análisis aparecen en la ventana de resultados (figura 2).

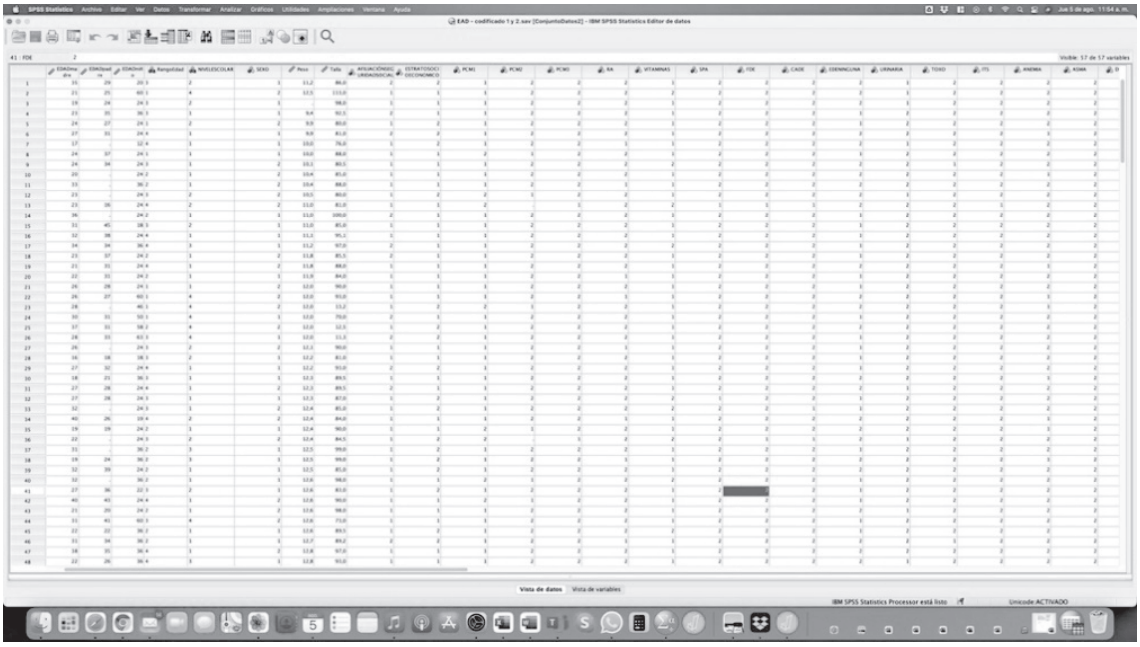


Figura 1. Pantalla con una base de datos en IBM-SPSS activa

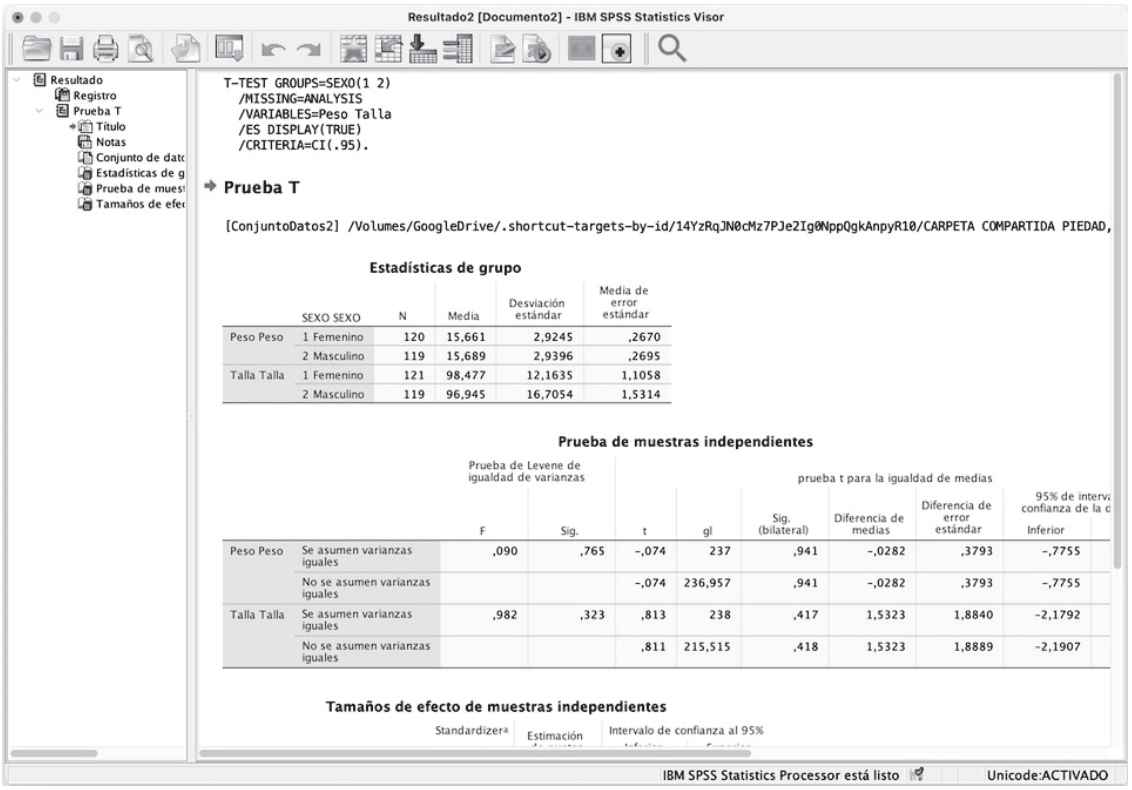
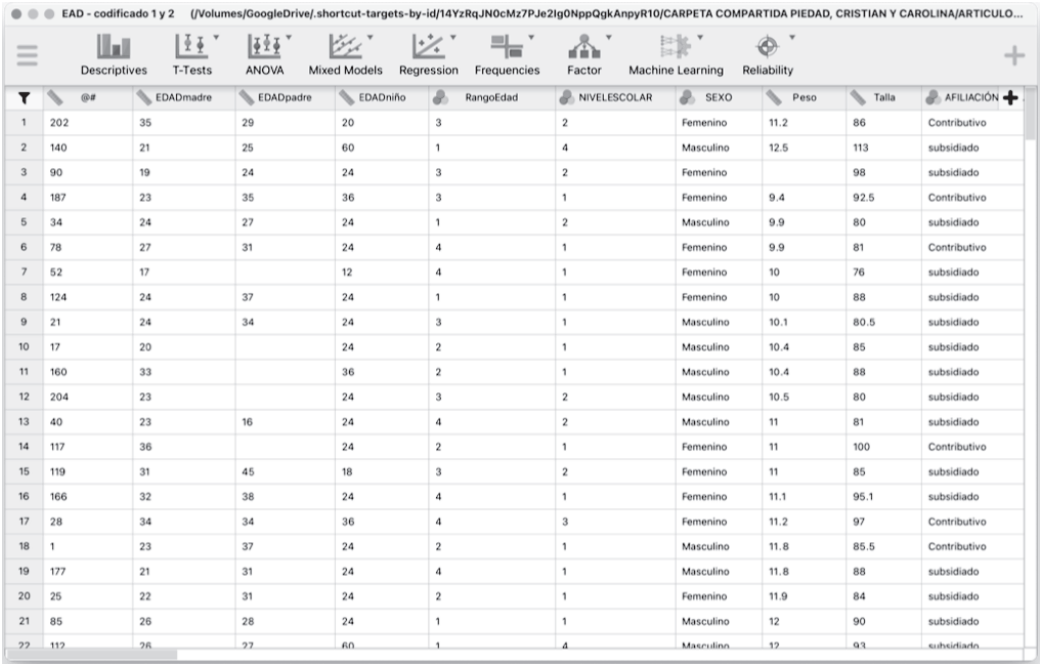


Figura 2. Ventana de resultados del IBM-SPSS

El segundo de los paquetes estadísticos que utilizamos es un paquete abierto multiplataforma para el procesamiento estadístico de datos que, además de tener interfaces intuitivas y fáciles de usar, no tiene costo: el JASP (Jeffrey's Amazing Statistics Program), desarrollado y actualizado de manera continua por un grupo de investigadores de la Universidad de Ámsterdam. Se puede descargar en línea y está disponible para Windows, MacOS X y Linux. En esta misma dirección se pueden descargar manuales y abundante documentación. El paquete y toda su documentación se encuentran en inglés. Por su facilidad, su integración y su enorme potencia, este paquete es altamente recomendable. En este libro se utiliza la versión 0.14.1, puesta en circulación en diciembre del 2020 y que corre bajo el sistema operativo MacOS Big Sur versión 14.5.

El JASP es bastante sencillo e intuitivo para su uso. En la misma ventana aparecen los accesos a la base de datos, los procedimientos de análisis y los resultados. Estos últimos se generan de inmediato. En la figura 3 aparece la forma en que se presenta la base de datos, una vez abierta en el programa. En la parte superior, aparecen los procedimientos disponibles; si se requirieran otros procedimientos, pruebe pulsar el signo "+", en la esquina superior derecha. Una vez se elija un procedimiento, la ventana cambiará a las opciones de procedimiento y los resultados, según aparece en la figura 4.



	@#	EDADmadre	EDADpadre	EDADniño	RangoEdad	NIVELESCOLAR	SEXO	Peso	Talla	AFILIACIÓN
1	202	35	29	20	3	2	Femenino	11.2	86	Contributivo
2	140	21	25	60	1	4	Masculino	12.5	113	subsidiado
3	90	19	24	24	3	2	Femenino		98	subsidiado
4	187	23	35	36	3	1	Femenino	9.4	92.5	Contributivo
5	34	24	27	24	1	2	Masculino	9.9	80	subsidiado
6	78	27	31	24	4	1	Femenino	9.9	81	Contributivo
7	52	17		12	4	1	Femenino	10	76	subsidiado
8	124	24	37	24	1	1	Femenino	10	88	subsidiado
9	21	24	34	24	3	1	Masculino	10.1	80.5	subsidiado
10	17	20		24	2	1	Masculino	10.4	85	subsidiado
11	160	33		36	2	1	Masculino	10.4	88	subsidiado
12	204	23		24	3	2	Masculino	10.5	80	subsidiado
13	40	23	16	24	4	2	Masculino	11	81	subsidiado
14	117	36		24	2	1	Femenino	11	100	Contributivo
15	119	31	45	18	3	2	Femenino	11	85	subsidiado
16	166	32	38	24	4	1	Femenino	11.1	95.1	subsidiado
17	28	34	34	36	4	3	Femenino	11.2	97	Contributivo
18	1	23	37	24	2	1	Masculino	11.8	85.5	Contributivo
19	177	21	31	24	4	1	Masculino	11.8	88	subsidiado
20	25	22	31	24	2	1	Femenino	11.9	84	subsidiado
21	85	26	28	24	1	1	Masculino	12	90	subsidiado
22	112	26	27	60	1	4	Masculino	12	93	subsidiado

Figura 3. Vista de la base de datos en JASP

Con frecuencia ilustraremos los resultados de un determinado procedimiento con una salida del paquete JASP o del IBM-SPSS. Al hacerlo, aportaremos información sobre la forma como, en cada uno de los dos paquetes estadísticos que manejamos, se puede ejecutar el procedimiento. Para denotar esta información, la expresaremos de la manera presentada en el recuadro 1. Obsérvese, por ejemplo, la instrucción para el cálculo de una prueba "*t*" para grupos independientes en el JASP (esta prueba se examinará, en detalle, en el capítulo 8).

Recuadro 1. Instrucciones para correr una prueba *t* sobre muestras independientes en JASP

/T-Test/Classical/ Independent Samples T-Test

En este punto, deben seleccionarse las variables dependientes (pueden ser varias) y pasarse a la lista "Variables" y la variable independiente, pasándola a la casilla "Grouping Variable" (debe ser una variable con solo dos valores)

Tests
- √ Student

Alt. Hypothesis
- √ Group 1 ≠ Group 2

Assumption Checks
- √ Normality
- √ Equality of variances

Additional Statistics
- √ Location parameter
  - √ Confidence interval [95,0 %]
- √ Effect Size
  - √ Cohen´s d
- √ Descriptives
- √ Descriptive plots
- Confidence interval [95,0 %]

Las instrucciones contenidas en el recuadro 1 deben entenderse de la siguiente forma:

> *En el menú raíz del JASP (/) debe seleccionarse T-Test, y allí, en la lista "Classical", el primero: "Independent Samples T-Test".*
>
> *Se anotan aquí las instrucciones a seguir en este punto y, más adelante, las especificaciones que se abren para este análisis. Cada vez que se anota el símbolo "√", significa que una casilla, con ese nombre, fue activada; posiblemente habrá otras que quedan sin activar. Por último, cuando encontramos valores entre corchetes, como [95,0 %], significa que puede modificar los valores allí contenidos.*

La forma en que esto es visible en el programa aparece en la figura 4. Para esta figura, se han seleccionado, en la base de datos, las variables "Peso" y "Talla" en la lista "Variables" y la variable "SEXO" en "Grouping Variable".

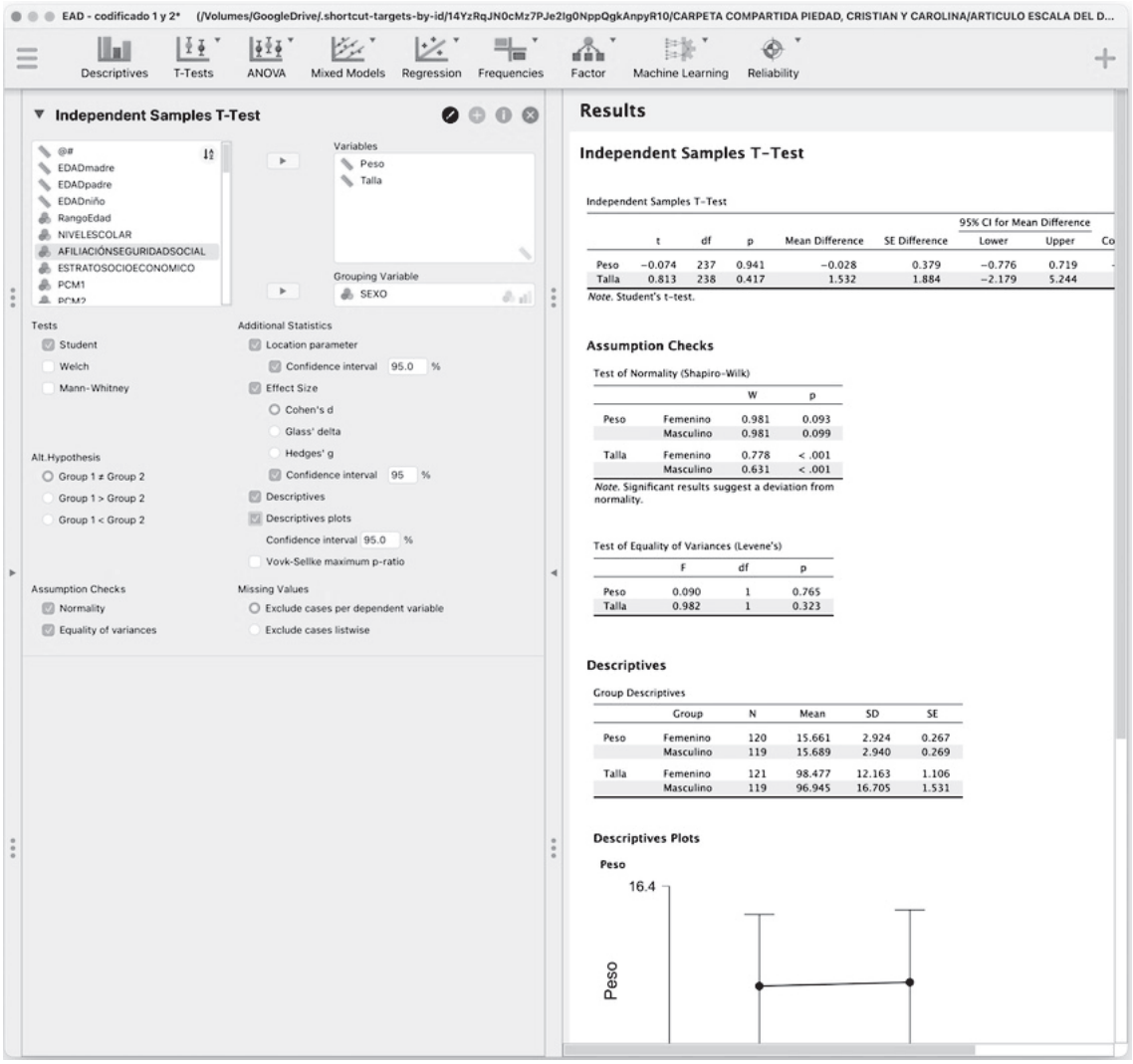


Figura 4. Ventana del JASP con las opciones seleccionadas en el recuadro 1 para una prueba *t* de Student

Como se observa en la figura 4, en la sección derecha aparecen directamente los resultados de los procedimientos solicitados. Esta característica del JASP facilita tomar decisiones sobre la marcha, teniendo en cuenta los resultados parciales. En este caso se observa, por ejemplo, que la variable "Talla" no cumple con el supuesto de normalidad, por lo que el investigador deberá seleccionar una prueba alternativa; aquí, la prueba *U* de Mann Whitney, presente en esta misma ventana bajo el título "Tests". Estas consideraciones se tratarán en el momento en el que examinemos la prueba *t*, en el capítulo 8.

En el IBM-SPSS, para esta misma prueba, las instrucciones quedarían como se indica en el recuadro 2.

Recuadro 2. Instrucciones para correr una prueba *t* sobre muestras independientes en IBM-SPSS

/Analizar/Comparar medias/Prueba T para muestras independientes…
En este punto deben seleccionarse las variables dependientes (pueden ser varias) y pasarse a la lista "Variables de prueba" y la variable independiente, pasándola a la casilla "Variable de agrupación".
En el botón "Definir grupos" deben anotarse los valores por comparar y pulsar el botón "Continuar"
√ Estimar tamaños del efecto
Pulsar el botón "Aceptar"

La anterior instrucción deberá entenderse de la siguiente forma:

> *En el menú raíz del SPSS (/) debe seleccionarse "Analizar" y, allí, "Comparar medias", a continuación, "Prueba T para muestras independientes…".*

Esto abre la ventana "Prueba T para muestras independientes" con listas de variables y botones que deberán ser manejados como se indica. Solo en las versiones más recientes del programa aparece el botón "Estimar tamaños del efecto"; es importante activarlo. En la figura 5 se muestra la apariencia de la ventana mencionada en el IBM-SPSS, versión 27.

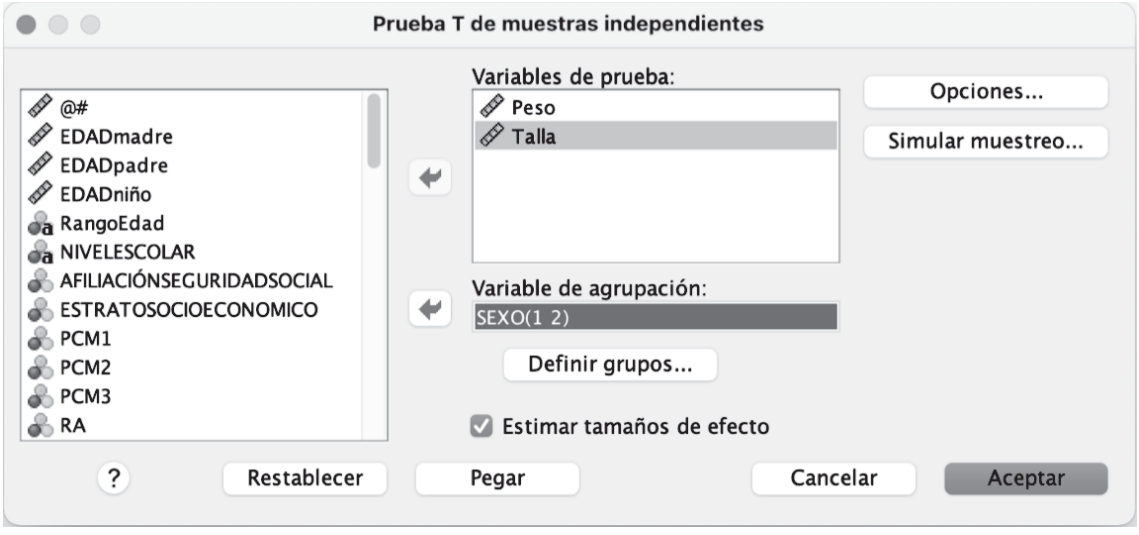


Figura 5. Ventana del IBM-SPSS con las opciones representadas en el recuadro 2 para una prueba *t* de Student

En este programa deberá pulsarse el botón "Aceptar" para que muestre los resultados del procedimiento en la ventana de resultados.

El plan de la obra

El libro inicia con una pequeña sección acerca de las definiciones iniciales y la confección de la base de datos. Los conceptos de población y muestra, parámetros y estadísticos, variable y valor y los diferentes tipos de escalas de medida, métricas y no métricas, son fundamentales en lo que sigue, por lo que deberemos hacer un repaso inicial al respecto. Hecho esto, podremos pasar a la confección de la base de datos. Esta deberá ser adecuadamente digitada y documentada. Parte de este proceso consiste en la depuración de la base para la corrección de los errores de digitación y la definición de los valores perdidos (*missing values*). Aunque tocaremos estos puntos en una sección inicial, no nos detendremos demasiado en ello, ya que muchas de las herramientas para la depuración serán tema de la parte subsecuente.

El siguiente momento del análisis consiste en realizar una adecuada descripción de los datos de los que disponemos. Ese será el tema de los capítulos 2 al 7. Para realizar la descripción de nuestros datos, se utiliza una gran variedad de estadísticos, a veces contenidos en tablas y gráficas que deben ser adecuadamente elaborados y presentados, a fin de dar una visión clara y sintética de lo que tenemos en nuestra base. Lamentablemente, los jóvenes investigadores, en su afán por llegar rápidamente a la prueba de sus hipótesis de investigación, con frecuencia pasan por alto una adecuada descripción de los datos. La experiencia indica que este grave error, con frecuencia, conduce a que todos los procesos deban repetirse muchas veces antes de llegar a soluciones finales.

Completada la descripción inicial de los datos, podemos pasar a la fase inferencial, en donde formulamos y probamos las hipótesis de la investigación. Este será el tema de los capítulos 8 al 12. Con frecuencia encontramos, en este momento, que las pruebas que queremos utilizar presentan uno o varios requerimientos, llamados "supuestos", que no siempre se cumplen para nuestros datos. ¿Qué hacer en este caso? Hay varias soluciones y la selección de la solución adecuada depende de la naturaleza de la investigación, del procedimiento y de nuestros datos. A veces es necesario volver atrás, hacer transformaciones de las variables y nuevas descripciones de ellas, o cambiar las pruebas que habíamos pensado por otras más resistentes a la violación de los supuestos, antes de llegar a examinar sus resultados.

Finalmente, nuestros resultados deben ser comunicados en un informe o, en el mejor de los casos, en un artículo científico. Para esto contamos con una serie de formatos en los que acostumbramos comunicar este tipo de datos. En particular, enfatizaremos en el estilo más usado para la comunicación de resultados estadísticos en la investigación social: el formato APA (denominado así por corresponder al adoptado por la American Psychological Association). En sus últimas versiones (6 y 7), el manual de estilo APA da una serie de indicaciones muy precisas acerca de las formas apropiadas para comunicar datos estadísticos, bien sea en el texto, o bien en gráficas o tablas. Al final de cada sección se pondrá información acerca de cómo utilizar este tipo de estilo para el procedimiento que acaba de ser expuesto.

Recomendaciones iniciales

La experiencia indica una serie de recomendaciones que valdría la pena tener en cuenta durante el procesamiento y análisis de datos:

- ¡Ojo con el almacenamiento de datos, programas y resultados! Iniciamos un proceso que se lleva a cabo, en su mayor parte, en computadores y, como decía un viejo amigo, "hay dos clases de usuarios de computadores: los que ya perdieron los datos y los que los van a perder". Esta frase hace explícita una experiencia común en la que, después de horas, o a veces días, perdemos todo el trabajo realizado y debemos volver a empezar. Esperamos que el lector no tenga que vivir esta experiencia o, por lo menos, no muchas veces, y que, si le ocurre, no tenga que devolverse demasiado. Para ello quisiéramos recomendar varias prácticas útiles.
  - Grabe la base de datos inicial, ojalá en la nube, con el nombre del proyecto en una carpeta dedicada a este, y grábela también en un dispositivo externo (una memoria USB, por ejemplo).
  - Inmediatamente lo haga, vuélvala a grabar con otro nombre e inicie el trabajo.
  - Cada transformación importante debe ser almacenada, ojalá en la nube, con diferentes nombres.
  - Si utiliza programas para el procesamiento, grábelos con nombres claros.
  - Grabe los resultados. En mi experiencia, prefiero trasladarlos a MS-Excel y grabarlos y graficarlos allí.
- Documente su base. Esto es importante a futuro. Utilice buenos nombres para las variables y etiquételas suficientemente. Si su base incluye ítems en un cuestionario, que la etiqueta de cada ítem sea el texto del ítem. Hágalo inmediatamente después de definir la variable, si lo deja para otro momento, puede olvidarlo. Etiquete los valores claramente. Cuando retome su base, unos años más adelante, debe poder entender qué hay allí.
- Sea parsimonioso en el análisis. Tómese su tiempo; conozca sus datos, deles vuelta una y otra vez; "amáselos" —si se permite la metáfora—. Mírelos por un lado y por el otro. Intente utilizar múltiples representaciones de ellos. Examínelos, primero globalmente, intentando comprenderlos, y después en sus diferencias entre grupos específicos. No le ponga límites a la estadística descriptiva. Conozca su base.
- Cuando llegue el momento de la estadística inferencial elija, inicialmente, la prueba más exigente: las pruebas paramétricas deben ser la primera opción. Corra las pruebas y verifique los supuestos. Si estos no se cumplen, intente evaluar las implicaciones del incumplimiento y solucionarlo con correcciones o trasformaciones. Si definitivamente no lo logra, y como último recurso, cambie de prueba por su equivalente no paramétrica. Recuerde incluir en el análisis alguna medida de tamaño del efecto.
- La regla de oro: sea ordenado. Diferencie lo importante de lo secundario y asegúrese de que puede encontrarlo y entenderlo más adelante. Puede llevar un diario de procesamiento en el que registre sus acciones. Al final, almacene todo en varias partes con etiquetas y nombres apropiados.

Capítulo 1

Para empezar

Definiciones iniciales

Estadística

Aunque el objeto específico de este libro no es, propiamente, la estadística, sino más bien el procesamiento cuantitativo de datos, la estadística nos da la gran mayoría de las técnicas y los conceptos que utilizaremos, por lo que deberemos profundizar un poco en su estudio.

Algo sobre su historia. En realidad, la estadística no es propiamente un desarrollo abstracto hecho por algunos matemáticos; por el contrario, tiene una historia que la vincula, desde sus inicios, con sus aplicaciones específicas.

La palabra *estadística* se deriva de la palabra italiana *statista*, 'persona que trata asuntos del Estado (*stato*)'. Originalmente llamada "la aritmética del Estado", se requirió para el cálculo de los ingresos y gastos del Estado, los impuestos y la posibilidad de financiación de guerras, por ejemplo.

De acuerdo con Aron y Aron (2002), el conocimiento estadístico proviene de situaciones muy diversas. Aunque la idea de recolectar datos inició por los requerimientos del gobernante, también se hizo por la necesidad de calcular la probabilidad de naufragios y de la piratería en los viajes marítimos de tiempos antiguos. Por su parte, los índices de mortalidad y los seguros de vida se iniciaron con el estudio de los cadáveres en las épocas de la peste negra. La predicción, por su lado, tiene su origen en la astronomía. La correlación proviene de la biología y, en particular, de la observación de los parecidos entre padres e hijos. La teoría de la probabilidad nace en el examen de las mesas de juego. La prueba de hipótesis surge en las destilerías y del estudio de las condiciones adecuadas para la agricultura. El análisis factorial se deriva de la psicología de la personalidad…

Estadística descriptiva e inferencial

En principio, debemos distinguir dos grandes ramas de la estadística: la estadística descriptiva y la inferencial.

La *estadística descriptiva* se encarga de la caracterización de un conjunto de datos y se utiliza para resumirlos y hacerlos comprensibles. Utiliza conceptos propios que logran resumir grandes

cantidades de información en unos pocos indicadores y que se pueden expresar en múltiples formatos, numéricos, tabulares o gráficos. Después de la introducción, dedicaremos los primeros cinco capítulos del libro, del capítulo 2 al 6, a la exposición de los elementos básicos de la estadística descriptiva.

Por su parte, la *estadística inferencial* se dedica al establecimiento de generalizaciones, o inferencias, de los datos obtenidos en pequeñas muestras de población a la población total. Esta rama nos permite examinar las posibilidades de que los resultados que obtenemos en pequeñas proporciones de la población puedan estar presentes en su totalidad. Es la parte más grande de la estadística y ocupa la última parte de este libro, desde el capítulo 7 hasta el 12.

Población y muestra

En la sección anterior mencionamos los conceptos de población y muestra, por lo que puede ser necesario definirlos de forma más precisa.

Una *población* es un conjunto completo de individuos o, en general, de elementos que poseen ciertas características comunes que el investigador desea estudiar. Aunque pareciera referirse a un conjunto de individuos humanos, esto no es necesariamente cierto. Dependiendo de nuestro interés, pueden ser ejemplares de una especie ubicada en un ecosistema, bacterias en la sangre de un sujeto, instituciones presentes en una localidad o artículos publicados sobre un tema, entre otros.

Por su parte, una *muestra* es una parte, más o menos pequeña, de elementos de la población que representan a la población misma. ¿En qué medida la representan? Eso será un tema posterior; desarrollaremos estos conceptos cuando introduzcamos los temas de la estadística inferencial.

Parámetros y estadísticos

Estos son conceptos muy relacionados con población y muestra. Un *parámetro* es un número calculado sobre los datos de una población, que cuantifica alguna característica de esa población. Por su parte, un *estadístico* es este mismo número, pero calculado a partir de los datos de una muestra, y que cuantifica una característica de esa muestra.

Es evidente que existe una relación clara entre parámetros y estadísticos. Una de las ramas de la estadística inferencial versa sobre las formas en que podemos estimar los *parámetros* de una población a partir de los *estadísticos* que obtengamos en una, o varias, muestras de ella. La presencia de rótulos diferentes para estos dos conceptos también es importante notarla.

Variables y valores

Una *variable* es cualquier característica de algún objeto, individuo o elemento que puede tomar diferentes valores, esto es, que varía, en una población dada. Ejemplos de variables en una población de estudiantes pueden ser su sexo, su nivel socioeconómico, sus evaluaciones en una materia dada, etc. Debe diferenciarse de una constante, que presenta siempre el mismo valor.

Por su parte, un *valor* es cada una de las posibilidades de variación de una variable. Si la variable es sexo, los valores pueden ser femenino o masculino, para no complicar demasiado las cosas (podrían incluirse diferentes tipos de valores intermedios, dependiendo de lo que entendamos por

"sexo"). Si la variable es nivel socioeconómico, los valores pueden ser "bajo", "medio", "alto", etc., dependiendo de cómo entendamos y volvamos operativa esa variable. Si la variable es el resultado de una evaluación en matemáticas, los valores pueden ser "E", "B", "A", "I", o también pueden estar codificados en una escala numérica de "1" a "10", por ejemplo, o representar el número de ejercicios correctamente resueltos en un examen, por ejemplo.

En este punto puede ser interesante distinguir las variables independientes de las variables dependientes. Esta distinción se presenta más bien en el caso del análisis de experimentos en los que se quiere contrastar una teoría que define una relación causal entre variables.

En general, entendemos la *variable dependiente* como aquella cuyos valores dependen de los valores tomados por otras variables, que consideraremos independientes. Las *variables independientes* son, en general, variables que manipulamos en los experimentos, mientras que las variables dependientes son aquellas que observamos para registrar si nuestra manipulación las afectó.

Supongamos, por ejemplo, que un investigador considera que el *logro de aprendizaje* en una materia depende, en alguna medida, de la *forma* en que el estudiante aborde el estudio de esa materia. El investigador quiere distinguir entre dos formas características: la primera, en la que el estudiante sigue, al pie de la letra, todas las actividades propuestas por el profesor sin separarse de ellas; y la segunda, en la que el estudiante aborda el estudio tomando en cuenta lo propuesto, pero complementándolo con otras actividades y fuentes. Tenemos acá dos variables y el planteamiento de una relación causal: el "logro de aprendizaje" es la variable dependiente, y "la forma en la que se estudia" es la variable independiente, con dos valores: 1) sujeta a las indicaciones del profesor y 2) autodirigida, y tenemos el planteamiento de una relación causal: si el estudiante estudia de una u otra forma, se afectará su logro de aprendizaje.

Niveles de medida

En este trabajo analizaremos datos que son producto de mediciones, por lo que resulta crucial manejar las escalas de medición.

Básicamente existen tres escalas de medición: las escalas nominales, las escalas ordinales y las escalas numéricas o métricas. Las diferencias entre ellas se refieren a la cantidad de atributos matemáticos que cada una posee.

La *escala nominal* se utiliza para variables de naturaleza más cualitativa que cuantitativa. Al utilizar una escala nominal, los valores de la variable resultan ser nombres, categorías o rótulos mutuamente excluyentes, sin que pueda establecerse una relación entre ellos. Ejemplo de variables nominales pueden ser el sexo (masculino, femenino), el color del pelo o de los ojos, la institución educativa en la que se está matriculado, la jornada escolar a la que asiste (mañana, tarde, única…), la religión que profesa, la tipología de familia, el municipio de nacimiento, etc.

Una propiedad fundamental de las escalas nominales es su equivalencia. Esto significa que todos los ejemplares de cada categoría son iguales desde el punto de vista de la variable. Todas las mujeres son equivalentes como mujeres; no hay ninguna que se distinga en ese sentido. Otra propiedad fundamental es la ausencia de una relación entre los valores: ningún valor puede ser considerado más o menos, que otro.

Puede ser importante distinguir entre dos tipos de escalas nominales: las dicotómicas y las politómicas. La diferencia estriba en el número de valores. En una *escala dicotómica*, también llamada *dummy*, la variable tiene solo dos valores, y representa la mínima cantidad de información posible (sí/no). En una *escala nominal politómica* se tienen más de dos valores. Ejemplo de este último tipo de escala puede ser la condición de empleo, definida con los valores "no activo", "desempleado", "empleado", "independiente" o "emprendedor". Esta distinción puede ser importante en la medida en que las variables dicotómicas presentan un procesamiento que puede ser mucho más sencillo que las politómicas.

En el siguiente nivel de medición aparecen las escalas ordinales. Una *escala ordinal* permite el ordenamiento, esto es, el establecimiento de una relación de orden (<) entre los valores. Para este tipo de variables podemos, entonces, identificar un valor que se considere "primero", un "segundo" y así sucesivamente. Ejemplos de este tipo de escalas son el nivel socioeconómico (alto, medio alto, medio…), las escalas evaluativas del tipo "muy bien", "bien", "regular", "mal" o (A++, A+, A, B, C…), el orden de llegada en una carrera, etc.

Una característica fundamental de las escalas ordinales es que, aunque ya tienen una propiedad de los números (el orden), no tienen mucho más y, en particular, no pueden establecer ni comparar la magnitud de las diferencias entre valores sucesivos. Esto significa que, para el caso del nivel socioeconómico, por ejemplo, la diferencia entre los niveles "muy alto" y "alto", en términos de ingresos, no es igual a la diferencia entre los niveles "bajo" y "muy bajo", aunque en los dos casos sean niveles sucesivos.

Algunos consideran que las variables nominales y ordinales tienen en común el no poder ser asimiladas a números, por lo que estos dos tipos de variables son agrupadas como *variables categóricas*. Así, hablaríamos de categorías nominales y categorías ordinales.

Por último, con el mayor nivel de medición, aparecen las escalas numéricas, métricas o propiamente cuantitativas. Una *escala numérica*, o *métrica*, hablando estrictamente, representa una magnitud o una cantidad de lo que se caracteriza. Este tipo de escalas contienen números con todas sus propiedades. Podemos conocer la diferencia entre dos valores, y esta es idéntica a la diferencia entre otros dos valores que presenten la misma distancia. Ejemplos de este tipo de escalas pueden ser el peso (medido en kilogramos), la altura (en centímetros), el tiempo de latencia de una respuesta (en segundos), el número de ejercicios correctamente resueltos en una prueba, etc.

Con frecuencia, en los libros de estadística se establecen diferencias entre dos tipos de escalas numéricas en relación con una propiedad final de los números: la presencia del cero (0) absoluto en la escala. Esto diferencia las escalas de intervalo y las escalas de razón. La *escala de intervalo,* o *escala intervalar*, tiene todas las propiedades de los números, excepto el "cero absoluto". El ejemplo clásico es la escala de temperatura en grados Celsius. En este caso, la diferencia entre 22 °C y 20 °C es exactamente igual a la diferencia entre 12 °C y 10 °C: esto es, 2 °C; sin embargo, el punto correspondiente a 0 °C no significa la ausencia total de la temperatura.

Este último punto es incorporado en las escalas numéricas de razón. En una *escala de razón* existe un valor "0", y este representa la ausencia total del atributo en cuestión. La escala Kelvin de temperatura es un ejemplo de este tipo de escala; el número de ejercicios correctamente resueltos en una prueba presentada, también lo es. La distinción entre las escalas de intervalo o de razón no

resulta demasiado importante en nuestro caso. Todos los procedimientos que pueden ser hechos en una escala de razón también pueden ser hechos en una escala de intervalo; la interpretación de los valores y las estimaciones sí que podría ser diferente.

La identificación del nivel de medición de una variable es crucial para la determinación la forma en que puede ser descrita, las gráficas que podemos emplear, el tipo de estadísticos que podemos utilizar y el tipo de pruebas que podemos usar con esa variable.

Variables continuas y discretas

Todavía tenemos que establecer una distinción entre las variables métricas: la que se establece entre variables continuas y discretas.

Una *variable continua* es aquella que, teóricamente, puede asumir un número infinito de valores entre dos unidades diferentes. Ejemplos de esta son las mediciones de tiempo, peso o longitud; entre un segundo y dos segundos pueden pensarse infinitos valores intermedios: 1,5 s, 1,51 s, 1,511 s…

Por su parte, una *variable discreta* es aquella para la cual no existen valores localizados entre puntos adyacentes de la escala. Ejemplos de esta pueden ser el número de ejercicios correctamente resueltos en una prueba: entre siete y ocho ejercicios no puede pensarse un número intermedio; el número de hijos en la familia es otro ejemplo de variable discreta.

Del procesamiento al reporte

En las siguientes secciones se describirán, de forma general, todos los pasos relacionados con el procesamiento y análisis de los datos, desde la construcción de la base hasta la expresión de los resultados en un artículo científico, pasando por la lectura de la base en los diferentes paquetes, la documentación de la base, la realización de modificaciones a las variables, la descripción de la información y la realización de inferencias.

Aunque pudiera parecer que este es un proceso secuencial en el que cada paso debe completarse para proceder al siguiente, no lo es. En efecto, cada resultado puede sugerir y mostrar nuevas alternativas que no habían sido consideradas inicialmente. El examen de una tabla de frecuencias puede mostrar que ciertos valores incluidos al principio en nuestros instrumentos no se requieren, o deben fundirse con otros. En estos casos, las variables deben modificarse o "recodificarse" y, de nuevo, describirse.

El proceso, en general, debe ser orientado por los resultados parciales que se van obteniendo. Con frecuencia, el analista deberá parar y devolverse a un paso anterior, redefinir una variable, ejecutar una transformación, documentarla y describirla antes de poder continuar. Con esta salvedad, describiremos los principales pasos.

La construcción de la base de datos

Iniciamos el procesamiento y análisis de los datos en la construcción de una base de datos. Una base de datos es, básicamente, la codificación en un formato de hoja electrónica, de los valores de las variables disponibles para un número determinado de casos. Identificamos los casos en filas y las variables en columnas, como se observa en la tabla 1.

Tabla 1. Modelo de base de datos en una hoja electrónica

ID del caso	Variable 1	Variable 2	…	Variable m
Caso 1	V_{11*}	V_{12}	..	V_{1m}
Caso 2	V_{21}	V_{22}	..	V_{2m}
Caso 3	V_{31}	V_{23}	…	V_{3m}
	…	…		…
	…	…		
Caso n	V_{n1}	V_{n2}	…	V_{nm}

Nota: * valor de la variable 1 en el caso 1, y así sucesivamente.

Para este caso, tendríamos una base con *n* casos y *m* variables. En general, la práctica más común es la digitación de los datos en una hoja electrónica del tipo MS-Excel, para su posterior lectura en el programa estadístico. En esta hoja se registran, en la primera fila, los nombres de las variables, utilizando algún tipo de código (por ejemplo: ID, I01, I02, …, In). Algunos prefieren incluir nombres que les recuerden el significado de la variable (p. ej., ID, sexo, grupo…)

Es importante anotar que, en la medida de lo posible, preferiblemente los valores V_{11} en la base deben ser valores numéricos. Esto no significa que la escala sea numérica; solo que el código que utilizamos será numérico. Por ejemplo, en vez de rotular "M" o "F" para la variable sexo, es preferible definir que "1" es "masculino" y "2" es "femenino", o algo similar. Para que este código sea interpretable, se deberá "documentar la base"; esto es, definir las etiquetas de las variables y las etiquetas de los valores de cada variable, cuando procede.

Con frecuencia encontramos que, en un conjunto de datos, no podemos incluir, para algún caso, algún valor, ya sea porque la información se perdió, no tiene sentido o no se comprende. Estos valores se conocen como *valores perdidos* (*missing values*, por su expresión en inglés). A este tipo de valores se les asigna un código especial (habitualmente " ", 9, 99, 999…), que deberá ser definido como tal, a fin de que no sea interpretado por el programa como un valor válido, sino que sea ignorado en el procesamiento.

En el caso de IBM-SPSS, existe también la posibilidad de digitar los datos directamente en el paquete estadístico. Esto puede hacerse en el menú /Archivo/Nuevo/Datos. Esto abrirá una matriz vacía de filas y columnas lista para ser llenada con la información.

Lectura de la base de datos

Los distintos paquetes que utilizamos tienen diferentes facilidades para cargar formatos de datos. El SPSS ha sido diseñado para trabajar con archivos de datos de formatos muy diversos, tales como hojas de cálculo de Excel, tablas de datos de Oracle, SQLServer, DB2, archivos ".csv" (*comma-separated values*), archivos de texto simples (.txt), archivos de datos SAS o Stata, etc. En el caso del JASP, las opciones son más limitadas; se pueden leer bases de datos en los formatos ".csv", ".txt" (archivos de texto), ".sav" (base de datos de SPSS), ".ods" (OpenDocument Spreadsheet) y, por supuesto, en su propio formato (.JASP).

En general, los programas hacen muy sencillo cargar una base de datos para su trabajo al marcar claramente los formatos en los que pueden hacerlo. Sugerimos partir de una base de datos previamente grabada en Excel. Los recuadros 3 y 4 indican cómo cargar bases de datos en los programas que usamos.

Recuadro 3. Cómo cargar una base de datos en JASP

=/=Open/Computer Browse

El símbolo "=" aparece en la esquina superior izquierda de la ventana del JASP. En este punto deberá navegar por los directorios de su computador hasta encontrar el archivo. Los formatos válidos aparecerán destacados, mientras que los inválidos aparecerán sombreados.

En este menú aparecerán otras opciones "Recent files", "OSF" y "Data Library". Recomendamos explorarlas.

Recuadro 4. Cómo cargar una base de datos en IBM-SPSS

/Archivo/Abrir/Datos

En este punto deberá navegar por los directorios de su computador hasta encontrar el archivo. Los formatos válidos aparecerán destacados, mientras que los inválidos aparecerán sombreados.

Documentación de la base de datos

Una vez se lee la base de datos en alguno de los paquetes estadísticos que utilizamos, se debe proceder a su documentación completa. Esto implica definir las etiquetas de los nombres de las variables, las etiquetas de los valores de las variables (cuando procede), los códigos para los valores perdidos, la definición del número de espacios y de decimales de cada variable y la definición del nivel de medición de la variable (nominal, ordinal o métrico). Para hacerlo, remitimos al lector a la documentación del *software* que esté usando.

Ya con la base leída y documentada, se pueden correr los primeros procedimientos de elaboración de las tablas de frecuencias para todas las variables. Estos primeros resultados permiten depurar la base y detectar algunos de los valores incorrectamente digitados, por estar fuera del rango posible para la variable. La base tiene que estar completamente depurada antes de proceder a su procesamiento. No debemos olvidar guardar la base documentada y depurada como un nuevo archivo y con un nuevo nombre.

Descripción

Una vez contamos con la primera versión documentada de la base de datos, podemos iniciar su descripción. Este proceso es fundamental y no debería ser ignorado, o minimizado —como habitualmente ocurre por parte de los investigadores sin experiencia—, en la ansiedad por llegar rápidamente a los resultados de las pruebas. Los elementos básicos de la estadística descriptiva aparecen, en este libro, entre los capítulos 2 y 6.

El examen descriptivo mostrará la necesidad de que algunas variables cambien. Algunos valores desaparecerán, algunas variables serán recodificadas en nuevas variables, se construirán variables nuevas que combinan otras, etc. En general, la base irá cambiando. Cada nueva variable —así como cada nueva versión de la variable— deberá ser adecuadamente documentada y descrita.

Inferencias

Una vez hayamos completado la descripción de los datos, se inicia la parte inferencial, en la que habitualmente elegimos y examinamos las pruebas de hipótesis adecuadas a nuestros objetivos y a nuestros datos. Los conceptos generales de la estadística inferencial se presentan en el capítulo 7 y las diferentes pruebas de hipótesis en el capítulo 9. A la selección inicial de las pruebas le sigue la verificación de los supuestos de la prueba, lo que significa, usualmente, aplicar otras pruebas para dicha verificación. Estos supuestos se presentan en el capítulo 9.

Si los supuestos de la prueba son satisfechos, podemos pasar a aplicarla y a examinar sus resultados. Si, por el contrario, no logramos satisfacerlos, aún es posible transformar las variables en nuevas formas que cumplan en mejor medida los supuestos. Algunos supuestos resultan críticos; en otros es posible elegir variaciones de la prueba particular. En los casos más extremos, nos veremos obligados a elegir una alternativa "libre de distribución" o "no paramétrica" a la prueba que habíamos elegido inicialmente. Por último, podremos ejecutar el programa y examinar los resultados de la prueba adecuada. Este proceso, para algunas de las pruebas más comunes, se presenta en este libro entre los capítulos 10 y 12.

Un punto adicional: cada vez más revistas científicas están requiriendo, además de los estadísticos y niveles de significación de las pruebas, una medida apropiada de "tamaño del efecto". Aportar este tipo de medidas puede ser un poco confuso puesto que, para cada prueba, existe un conjunto diferente de medidas de tamaño del efecto que resultan apropiadas. Para completar la dificultad, muchos paquetes de *software* no incluyen el cálculo de este tipo de medidas; el JASP, en general, incluye este tipo de medidas, pero el IBM-SPSS solo las tiene para algunos procedimientos. La presentación de cada medida de tamaño de efecto será hecha en cada prueba de hipótesis.

La expresión de resultados: ¿texto, tablas o gráficas?

Finalmente, los resultados obtenidos deberán ser expresados en un informe técnico de investigación o, en el mejor de los casos, en un artículo científico que se publicará en alguna revista. Esto exige adecuarse y adaptarse a las particulares formas de comunicación científica. En cada proceso y prueba, hemos tratado de incluir un apartado final acerca de cómo deben comunicarse los resultados.

Primero, una anotación general sobre el número de decimales y el punto o coma decimal. Como regla general, en inglés se utiliza el punto decimal (p. ej., 3.14), mientras que en español utilizamos la coma (p. ej., 3,14) para representar la separación entre los enteros y los decimales. En ningún caso se utiliza el punto para la separación de miles, ya que esto induciría a confusiones.

El número de decimales, por su parte, está claramente definido en el formato APA. La regla es simple: siempre se utilizan dos posiciones decimales después del punto, o coma decimal (p. ej., *M*=24,18), excepción hecha de la expresión del nivel de significancia, o "valor p", que siempre se representa con tres posiciones decimales y sin el cero correspondiente a los enteros (p. ej., *p*=,043).

Existen tres formas para expresar los resultados cuantitativos en un informe o en un artículo: en el texto, en tablas o en gráficas. En general, si presentamos estadísticas en una tabla o en una gráfica, no es necesario repetirlo en el texto (American Psychological Association [APA], 2010).

Para decidir cuál es la mejor forma de hacerlo, la sexta edición del manual de estilo APA recomienda seguir una regla general, que parafraseamos a continuación:

- Si necesita presentar tres números, o menos, inténtelo en una oración.
- Si necesita presentar entre cuatro y veinte números, considere usar una tabla bien preparada.
- Si tiene más de veinte números, a menudo una gráfica resulta más útil que una tabla (APA, 2010).

Aunque ese criterio de decisión puede ser útil, no es determinante. Es importante que examinemos cuidadosamente las ideas centrales que quisiéramos enfatizar y omitamos detalles innecesarios o distractores. Algún medio permitirá expresar con mayor claridad esa idea.

Con frecuencia algunas gráficas, cuidadosamente planeadas, pueden transmitir con mayor claridad y contundencia datos sobre diferencias. Sin embargo, en general, las gráficas representan dificultades para los editores de las revistas, por lo que se recomienda minimizar su uso o, al menos, restringirlo a lo más importante. El formato gráfico en el que se presenten (.jpg, .tif…) y el uso de color dependen de la revista específica. Recientemente, con la generalización de la publicación en formato electrónico, las revistas han incluido color en sus gráficas, algo que antes no era posible.

Las tablas, por su parte, pueden expresar información muy precisa, siempre que estén bien diseñadas y, en general, no plantean tantos problemas a los editores. Es bastante frecuente encontrar tablas, incluso de gran tamaño, en artículos científicos. En el formato APA, las tablas solo contienen líneas horizontales que las enmarcan, y que separan los títulos de las columnas. Un ejemplo de una tabla adecuadamente diseñada en formato APA puede ser el que se observa en la tabla 2.

Tabla 2. Ejemplo de una tabla en formato APA

Tabla x. Resultados de las pruebas *t* para grupos independientes

Prueba	Grupo	M	DE	t	gl	p	*d* de Cohen
Pretest	Experimental	78,28	20,88	-0,10	48	,986	-0,01
	Control	78,37	18,97				
Postest	Experimental	102,32	31,95	21,00	48	,014*	0,72
	Control	81,32	26,13				

***p<,05**

Tanto tablas como gráficas deberán estar adecuadamente tituladas y numeradas. Obsérvese que en esta tabla hemos incluido toda la información relevante: estadísticos descriptivos (*M*, *DE*), resultados de la prueba de hipótesis (*t, gl, p*) y una medida del tamaño del efecto (*d* de Cohen).

La expresión de los resultados en el texto es la forma más fácil de hacerlo, desde el punto de vista del editor, pero es la que más trabajo exige de los autores. A veces es un verdadero rompecabezas expresar diferentes resultados de pruebas en un párrafo manteniendo la claridad en el mensaje. Sin embargo, cuando es posible, actualmente se privilegia este tipo de expresión de los resultados.

Para la expresión de los resultados en texto, la séptima edición del manual de la APA (2020) recomienda presentar todos los estadísticos disponibles, incluyendo medidas de tendencia central y dispersión, estadísticos de prueba, grados de libertad, niveles de significación y medidas de tamaño del efecto, siguiendo un formato que resulta propio de cada prueba. Por poner un ejemplo, para la presentación de los resultados de una prueba *t* de Student sobre grupos independientes, se deberán incluir, al menos: 1) medias y desviaciones estándar de la variable dependiente en cada grupo, 2) los grados de libertad de prueba, 3) la medida del estadístico de prueba *t*, 4) la medida de valor "p", o nivel de significación, y 5) una medida apropiada de tamaño del efecto que, en esta prueba, es la *d* de Cohen. Siguiendo el formato adecuado para la prueba *t*, un texto que presenta los resultados del postest, en la tabla 2, puede quedar como se observa en la figura 6.

> Las medias (con desviaciones estándar entre paréntesis) para la prueba en los dos grupos fueron de 81,31(26,12) para el grupo de control y de 102,32(31,94) para el grupo experimental. Los resultados de la aplicación de la prueba t de Student sobre grupos independientes muestran que la diferencia entre estos grupos es significativa a nivel de ,05 con un tamaño del efecto mediano $t(48)=2,54$ $p =,014$ $d = 0,72$.

Figura 6. Ejemplo de texto que presenta los resultados de una prueba *t*

Un punto final. Las formas en que se exponen los resultados de una prueba estadística en texto dependen de la prueba y se mostrarán en el momento en el que presentemos la prueba misma. Debe recordarse que, de acuerdo con APA, además de los niveles de significación es altamente recomendable presentar una medida apropiada de tamaño del efecto que, de igual forma, dependerá de la prueba específica. El concepto de tamaño del efecto se estudiará, en el capítulo 8.

Capítulo 2

Frecuencias, percentiles y representaciones gráficas

Presentación

En esta parte, correspondiente a la estadística descriptiva, nos ocuparemos de las formas y recursos de los que disponemos para describir, de manera ordenada y comprensiva, un conjunto de datos, independientemente del número de observaciones que contenga. Para ello, el primer y principal recurso del que disponemos y que resulta útil, independientemente del nivel de medición de la variable, es la tabla de frecuencias.

Presentadas las tablas de frecuencias, podremos examinar las formas de representar los datos en gráficas, hasta llegar al histograma, como forma de incorporar la distribución de variables numéricas. Completado esto, podremos estudiar las diferentes formas de caracterizar las distribuciones de frecuencias y el uso de criterios numéricos para compararlas con la distribución normal. Finalmente, podremos estudiar los puntos y grupos percentiles.

Frecuencias y distribuciones

Frecuencias simples. Tablas y representaciones gráficas

A continuación, aparece el listado de valoraciones hechas por diferentes maestros de Matemáticas en octavo grado, acerca del rendimiento de sus estudiantes en la materia. Los datos están dados en una escala ordinal de cuatro puntos, del 1 al 4, en la que "1" representa un rendimiento "deficiente"; "2", "aceptable"; "3", "superior", y "4", "excelente". En total se presentan 231 datos. Observe el listado de la figura 7.

1, 2, 1, 1, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 3, 2, 2, 2, 2, 1, 2, 1, 1, 1, 3, 3, 1, 2, 2, 2, 2,
2, 2, 3, 1, 3, 3, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 3, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 3, 2, 2, 2, 2, 1, 1, 2, 1, 2, 1, 2,
2, 2, 2, 2, 2, 2, 1, 3, 2, 2, 1, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 3, 2, 1, 1, 1, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2,
3, 4, 1, 3, 2, 3, 2, 2, 1, 1, 2, 2, 2, 1, 1, 4, 1, 2, 2, 3, 3, 2, 2, 3, 2, 1, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 4, 2, 2, 2, 3,
3, 3, 3, 1, 1, 2, 1, 1, 2, 2, 2, 3, 2, 2, 2, 2, 1, 2, 2, 3, 1, 3, 2, 4, 1, 3, 4, 2, 1, 3, 2, 4, 2, 2, 1, 4, 2, 2, 2, 3, 3, 3,
2, 2, 1, 3, 2, 2, 2, 1, 2, 2, 2, 4, 2, 2, 2, 2, 2, 1, 4, 1, 2

Figura 7. Listado con 231 valoraciones de un maestro de matemáticas en una escala que va de "1" a "4"

A partir de esta observación, ¿podría usted indicar cómo les fue a los estudiantes con respecto a la valoración que hacen sus docentes en el curso? ¿Bien? ¿Mal?

Como se observa, es difícil hacer siquiera una descripción general de los datos. Por inspección visual notamos que el número "4" es más escaso, y que el número "2" parece ser muy frecuente, pero poco más podemos decir. La dificultad de la tarea se incrementa de forma considerable en la medida en que aumenta el número de valores posibles (por ejemplo, que no sean 4, sino 10) o el número de observaciones (imagine no 231 observaciones, sino 1000, o 5000). La inspección visual le puede tomar un tiempo considerable y al terminar apenas se tiene una idea imprecisa de la tendencia general.

Una solución para una primera descripción de lo que hay en la base es elaborar una *tabla de frecuencias*. Esto es, una tabla que muestre cuántos alumnos obtuvieron cada uno de los valores posibles o, lo que es lo mismo, con qué *frecuencia* se obtiene cada valor.

Para hacerlo, hace unos cincuenta años hubiéramos tenido que seguir un largo proceso, que inicia con escribir el listado de todos los valores posibles (en este caso, del 1 a 4), y recorrer el listado, de principio a fin, poniendo una marca (un *palito*) al frente de cada valor a medida que aparece, y tachando el valor, para no desorientarse en un proceso tan largo como tedioso. Al final, contaríamos el número de marcas al frente de cada valor para obtener su frecuencia. Un proceso dispendioso que, afortunadamente, hoy en día no tenemos que seguir.

Para calcular una tabla de frecuencias simples en JASP, puede procederse de la forma expresada en el recuadro 5. El programa solo generará tablas de frecuencias para variables categóricas (nominales u ordinales) y presentará, además de las tablas de frecuencias, algunos datos descriptivos que resultan útiles (valores válidos, *missing*, mínimo, máximo) y algunos estadísticos, que serán estudiados más adelante (medias y desviaciones estándar). Cuando se hace en el IBM-SPSS, debe procederse como se muestra en el recuadro 6.

Recuadro 5. Instrucciones para solicitar una tabla de frecuencias en JASP

/Descriptives
En este punto se arrastran las variables que deseamos describir a la lista "Variables"
√ Frecuency tables (nominal and ordinal variables)

Recuadro 6. Instrucciones para solicitar una tabla de frecuencias en IBM-SPSS

/Analizar/Estadísticos descriptivos/Frecuencias…
En este punto se arrastran las variables que deseamos describir a la lista "Variables"
√ Mostrar tablas de frecuencias

El IBM-SPSS no tiene restricciones al hacer tablas de frecuencias de variables numéricas y continuas. Al hacerlo en el programa IBM-SPSS, se arroja una tabla como la 3.

Tabla 3. Tabla de frecuencias de la variable "evaluación del maestro de Matemáticas"

Evaluación del maestro de Matemáticas					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Deficiente	59	25,5	25,5	25,5
	2 Aceptable	132	57,1	57,1	82,7
	3 Superior	31	13,4	13,4	96,1
	4 Excelente	9	3,9	3,9	100,0
	Total	231	100,0	100,0	

Primero, observe que en la última fila se representa el número total de casos para nuestro ejemplo: 231. Este valor representa el 100,0 % de los casos totales y el 100,0 % de los casos válidos, ya que no hay datos perdidos, o *missing*.

En esta tabla aparecen, además de la columna de valores, cuatro columnas más: "Frecuencia", que representa el conteo simple de cada valor; "Porcentaje", que representa su porcentaje dentro del total de casos; "Porcentaje válido", que solo diferirá del anterior si tenemos valores perdidos (y por tanto no válidos); y "Porcentaje acumulado", que representa el porcentaje de la suma acumulada de los valores ya listados.

Ahora sí, podemos interpretar el resultado. De acuerdo con lo encontrado, una gran mayoría de los estudiantes (132), que representa el 57,1 % de los casos, es valorado por su maestro como "aceptable"; en orden de frecuencia, la segunda categoría más frecuente es "deficiente", en la que se ubican 59 alumnos (25,5 %).

Como se observa, la tabla de frecuencias ofrece gran cantidad de información importante. El uso de la columna correspondiente al porcentaje acumulado nos permite producir nuevos datos; podríamos también decir, por ejemplo, que aproximadamente tres de cada cuatro estudiantes lograron notas aprobatorias (100 %–25,5 % = 74,5 %), o que solo el 17,3 % (100 %–82,7 %) tienen valoraciones más altas que lo apenas aceptable. El foco de atención depende de aquello que queremos enfatizar.

Los datos contenidos en las tablas de frecuencias pueden ser representados gráficamente de varias formas. Las más utilizadas para datos categoriales, nominales u ordinales de pocos valores son las *gráficas de barras* y las *gráficas circulares*, también llamadas de "tortas" o de "pie". Cada una de estas enfatiza diferentes aspectos de la información. Observe las gráficas de la figura 8, generadas por el SPSS en la opción "Gráficas", del procedimiento "Frecuencias" que estamos estudiando.

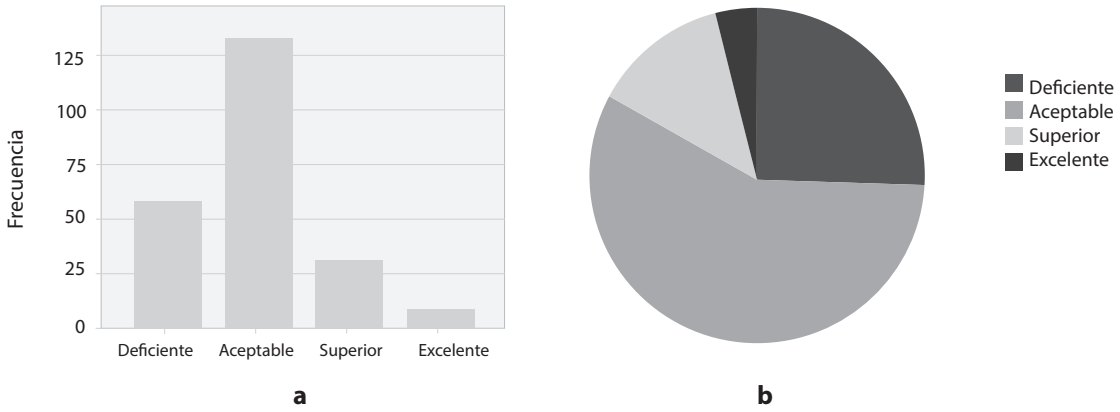


Figura 8. Representaciones gráficas de la variable "evaluación del maestro de Matemáticas"
Nota: a) gráfico de barras; b) gráfico circular.

A la izquierda, en la figura 8a, tenemos una gráfica de barras. A partir de esta, es clara la comparación de las frecuencias de los diferentes valores. Es fácil ver, por ejemplo, que la cantidad de "aceptables" casi duplica al siguiente valor, en orden de frecuencia ("deficientes"), y que los que siguen van en orden descendente. Se ve que la cantidad de alumnos rotulados con rendimiento "excelente" es la más pequeña.

En la figura 8b encontramos una gráfica circular. En este caso, hay un énfasis, que no se hacía en el anterior, en la *proporción* que representa cada frecuencia dentro del total. Así, es posible saber, no solo que el valor "aceptable" es el más frecuente, sino que por sí mismo representa más de la mitad de los datos totales. También se puede saber, con un golpe de vista, que la proporción de alumnos rotulados con rendimiento deficiente es casi del 25 %, o uno de cada cuatro. ¿Qué gráfica elegir? Depende de la situación y de lo que el investigador quiera enfatizar: valores o proporciones.

Puede ser interesante anotar que la elaboración de tablas de frecuencias no es una actividad exclusiva hecha sobre bases de datos. Muchas técnicas de observación, ya sea *in situ* o a partir de un registro de video, requieren de la elaboración de una tabla de frecuencias que permita, por ejemplo, registrar comportamientos en un salón de clases (p. ej., número de veces que un estudiante desatiende al profesor para interactuar con compañeros, o número de preguntas hechas al profesor por los estudiantes).

Frecuencias agrupadas

En el ejemplo anterior, construimos tablas y gráficas que representaban frecuencias simples de una variable ordinal de cuatro puntos. Ahora, cuando la variable es propiamente numérica, y muy especialmente cuando presenta muchos valores, surgen algunas dificultades y se abren nuevas posibilidades de representación.

Consideremos ahora un segundo ejemplo. Como parte de esta misma base de datos, disponemos de información sobre los resultados obtenidos por los 231 alumnos frente a una prueba, muy

utilizada por nuestro grupo de investigación, conocida como la prueba EFT. Esta se usa en la medición de cierta capacidad perceptual visual conocida como "capacidad de reestructuración cognitiva", que se asocia con un estilo cognitivo. La prueba produce resultados que pueden variar entre 0 y 50 puntos, correspondientes a la misma cantidad de ejercicios correctamente resueltos en un tiempo determinado.[1]

Los resultados aparecen en la tabla 4. Como se observa, en este caso la tabla es bastante más grande, por lo que tuvimos que editarla, suprimir columnas y acomodarla en el espacio de una página. En este caso, aunque la tabla de frecuencias nos da mucha información, no resulta tan fácil de utilizar por la presencia de un número grande de valores.

Una solución a este problema consiste en agrupar valores en intervalos; esto es, formar intervalos iguales que agrupen los datos. Esto se conoce como *frecuencias agrupadas*. En el caso de nuestro ejemplo, tenemos hasta 50 valores posibles dentro de nuestra prueba, por lo que podemos formar intervalos de variada longitud. Podemos formar, por ejemplo, veinticinco intervalos, cada uno con una longitud de dos puntos o diez intervalos con una longitud de cinco puntos o cinco intervalos, cada uno con una longitud de diez puntos.

Para hacerlo, los programas estadísticos brindan varias posibilidades muy útiles. El IBM-SPSS ofrece, entre otras, las opciones de "Recodificar" (/Transformar/Recodificar en distintas variables…)[2] y "Agrupación visual" (/Transformar/Agrupación visual…). En el primer caso, se debe definir una nueva variable y definir cada intervalo por separado. En el segundo, después de seleccionar la variable original, se pueden especificar el punto de inicio y el número de intervalos. Vale la pena explorar estas dos posibilidades.

Tabla 4. Tabla de frecuencias de la variable "Puntaje EFT"

	Frecuencia	Porcentaje	Porcentaje acumulado		Frecuencia	Porcentaje	Porcentaje acumulado
0	3	1,3	1,3	26	11	4,8	58,9
3	2	0,9	2,2	27	9	3,9	62,8
5	3	1,3	3,5	28	9	3,9	66,7
6	3	1,3	4,8	29	5	2,2	68,8
7	2	0,9	5,6	30	10	4,3	73,2
8	3	1,3	6,9	31	5	2,2	75,3
10	3	1,3	8,2	32	7	3,0	78,4
11	2	0,9	9,1	33	7	3,0	81,4
12	5	2,2	11,3	34	10	4,3	85,7
13	2	0,9	12,1	35	5	2,2	87,9
14	3	1,3	13,4	36	2	0,9	88,7

1 La denominación EFT proviene del inglés *Embedded Figures Test*. Para mayor información de esta prueba, puede consultar Hederich (2004).

2 Este representa el camino por seguir desde el directorio raíz del programa.

	Frecuencia	Porcentaje	Porcentaje acumulado		Frecuencia	Porcentaje	Porcentaje acumulado
15	5	2,2	15,6	37	5	2,2	90,9
16	5	2,2	17,7	38	3	1,3	92,2
17	8	3,5	21,2	39	2	0,9	93,1
18	5	2,2	23,4	40	6	2,6	95,7
19	6	2,6	26,0	41	3	1,3	97,0
20	9	3,9	29,9	43	1	0,4	97,4
21	6	2,6	32,5	44	1	0,4	97,8
22	13	5,6	38,1	45	1	0,4	98,3
23	14	6,1	44,2	46	2	0,9	99,1
24	15	6,5	50,6	47	1	0,4	99,6
25	8	3,5	54,1	48	1	0,4	100,0
				Total	231	100	

En nuestro caso, con propósitos pedagógicos, hemos definido dos nuevas variables que contienen los datos agrupados de la prueba EFT. Por un lado, definimos una variable con diez intervalos, cada uno con una longitud de cinco puntos; por otro lado, definimos otra variable con cinco intervalos, cada uno con una longitud de diez puntos. Las tablas 5 y 6 representan la distribución de frecuencias de estas dos nuevas variables.

Tabla 5. Frecuencias agrupadas de la variable EFT en diez grupos

EFT10 EFT de 10 puntos					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1,00 [0, 5)	5	2,2	2,2	2,2
	2,00 [5,10)	11	4,8	4,8	6,9
	3,00 [10, 15)	15	6,5	6,5	13,4
	4,00 [15, 20)	29	12,6	12,6	26,0
	5,00 [20, 25)	57	24,7	24,7	50,6
	6,00 [25, 30)	42	18,2	18,2	68,8
	7,00 [30, 35)	39	16,9	16,9	85,7
	8,00 [35, 40)	17	7,4	7,4	93,1
	9,00 [40, 45)	11	4,8	4,8	97,8
	10,00 [45, 50)	5	2,2	2,2	100,0
	Total	231	100,0	100,0	

Tabla 6. Frecuencias agrupadas de la variable EFT en cinco grupos

EFT5 EFT de cinco puntos					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1,00 [0, 10)	16	6,9	6,9	6,9
	2,00 [10, 20)	44	19,0	19,0	26,0
	3,00 [20, 30)	99	42,9	42,9	68,8
	4,00 [30, 40)	56	24,2	24,2	93,1
	5,00 [40, 50]	16	6,9	6,9	100,0
	Total	231	100,0	100,0	

Es importante observar que, para definir cada clase, o intervalo, tomamos aquellos valores iguales o mayores a un límite inferior, y los estrictamente menores al límite superior. Ese es el sentido de denominar a los intervalos de la forma [a, b); este intervalo queda definido por todos los números mayores o iguales al número a, y menores al número b; esto es, todos los x, tales que $a \leq x < b$.

En la figura 9 aparecen las gráficas de barras de cada una de estas nuevas variables. Se han dispuesto en esta forma para facilitar la comparación.

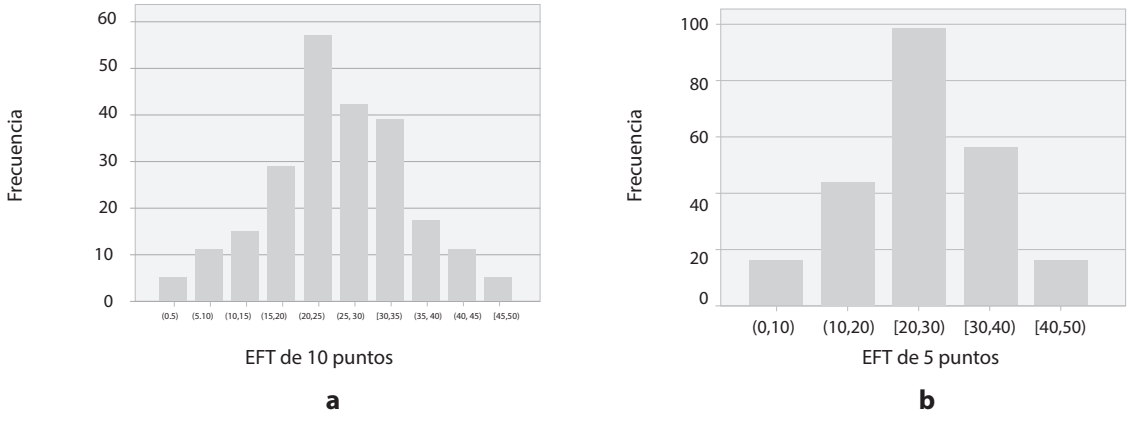


Figura 9. Gráfica de barras de frecuencias agrupadas de la variable EFT

Nota: a) frecuencias agrupadas en diez intervalos; b) frecuencias agrupadas en cinco intervalos.

Como se observa, las formas de las gráficas de barras son bastante similares. Las barras en el área central son mucho más altas, mientras que hacia los extremos decaen. Este es un comportamiento típico que más adelante examinaremos con detalle.

Para confeccionar una tabla de frecuencias agrupadas se debe considerar una serie de principios:

- Es ideal tener entre 5 y 15 intervalos como máximo. Un número menor que 5 implica perder demasiada información, y un número mayor que 15 mantiene el problema que tratábamos de solucionar al agrupar las frecuencias.

- Es importante trabajar con tamaños en cada intervalo que nos resulten cómodos: 2, 3, 5 y 10 son valores usuales, o los múltiplos de 10.
- Es muy importante que todos los intervalos tengan igual longitud; de lo contrario, podríamos estar deformando la gráfica.

Más gráficas para representar frecuencias

Existe un tipo especial de gráfica, similar a la gráfica de barras, particularmente apropiada para graficar frecuencias en variables numéricas: el *histograma.* Se trata de una gráfica similar a la de barras, ya que, en esta, la altura de la barra representa la frecuencia que le corresponde al intervalo. Sin embargo, existe una diferencia fundamental entre la gráfica de barras y el histograma, y es que el último trata los valores de la variable, dispuestos en el eje X, como valores *propiamente numéricos*, mientras que la gráfica de barras los trata como rótulos. Esto tendrá consecuencias importantes en la forma de la curva.

La figura 10a representa el histograma de la variable EFT, tal y como lo produce el SPSS, y la figura 10b representa la gráfica de barras de la misma variable. ¿Observa alguna diferencia?

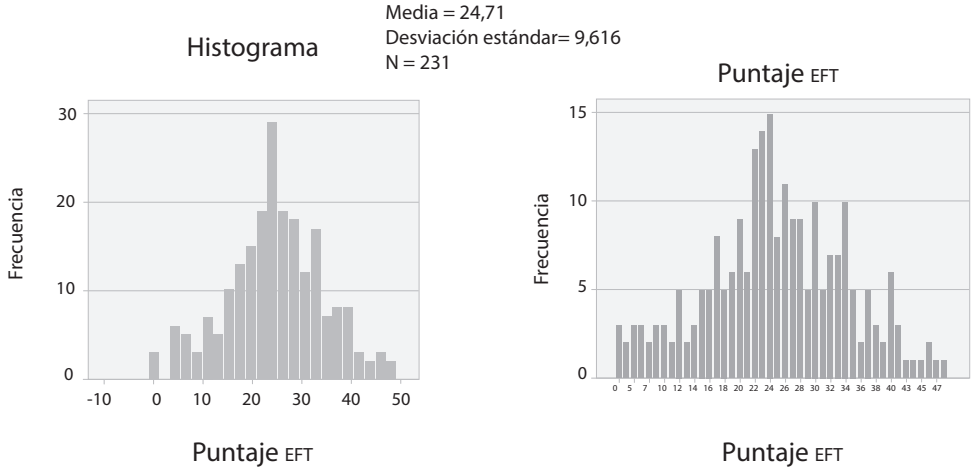


Figura 10. Dos representaciones gráficas de la variable EFT

Nota: a) histograma; b) gráfica de barras.

Existen varias diferencias importantes. Primero, en la gráfica 10a se ha agrupado la variable de puntaje EFT en 25 intervalos, cada uno con longitud de dos puntos, lo que da una visión más compacta y eleva la escala del eje vertical, al doble de la inicial. Esto hace que, en comparación con la gráfica de barras, el histograma parezca un poco más suave, con menores picos.

Segundo, y esto es importante, debe observarse un espacio vacío entre la primera y la segunda columna del histograma, que no aparece en las barras. Esto ocurre porque en el histograma se consideran todos los valores en la escala numérica, haya o no haya datos, mientras que en las barras solo se consideran los valores donde efectivamente hay datos. Por esta razón, en la gráfica de barras los números del eje X aparecen con saltos irregulares (0, 5, 7, 10, 12…), mientras que en el histograma la variable numérica, en el eje X, está dispuesta uniformemente (0, 10, 20…). Para subrayar las diferencias entre estas dos representaciones, en las gráficas de barras las barras aparecen separadas, aunque sea levemente, mientras que en los histogramas permanecen unidas.

Por último, vale la pena observar que en el histograma generado por el SPSS se anotan los datos de tendencia central de la variable: media, desviación estándar y tamaño de muestra. Este tipo de medidas se estudiarán en el siguiente capítulo. Esto solo se puede hacer porque la variable es propiamente numérica.

No en todos los programas se generan los histogramas de la misma forma. Obsérvese en la figura 11 el histograma generado por el programa JASP para la variable EFT en su forma original de 50 puntos.

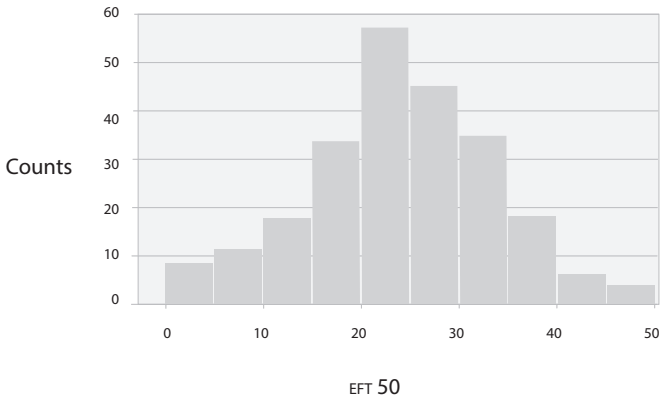


Figura 11. Histograma de la variable EFT producido por el JASP

En este histograma no se hicieron frecuencias agrupadas en 25 puntos, sino en 10 puntos. La diferencia, en realidad, no es muy importante.

Existe otra gráfica de uso común para representar frecuencias: el llamado *polígono de frecuencias*. Consiste en que, a partir de un histograma, se toman los puntos medios de la parte superior de las barras y se traza una gráfica de líneas, sombreando el área resultante. Observe el polígono de frecuencias de la figura 12, para la misma variable que hemos estudiado. Para generar un polígono de frecuencias, en el SPSS, siga este camino: /Gráficos/Generador de gráficos/Histograma.

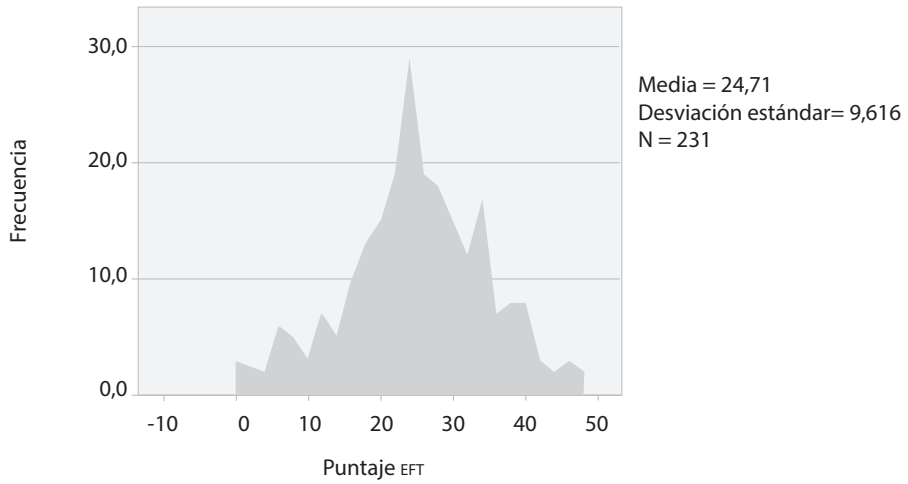


Figura 12. Polígono de frecuencias de EFT

Existe otro tipo de gráfica con capacidad para representar frecuencias de una variable. Se conoce como *diagrama de tallo y hojas*. La figura 13 muestra un diagrama de tallo y hojas de la misma variable, tal y como es generado por el SPSS. Este tipo de gráficas no está disponible en JASP.

Puntaje EFT Gráfico de tallo y hojas
Frecuencia Stem & Hoja
3,00 Extremes (=<0)
2,00 0 . 33
11,00 0 . 55566677888
15,00 1 . 000112222233444
29,00 1 . 55555666667777777788888999999
57,00 2 . 000000000111111222222222222333333333333333444444444444444
42,00 2 . 555555556666666666777777777888888888999999
39,00 3 . 000000000011111222222233333334444444444
17,00 3 . 55555667777788899
11,00 4 . 00000011134
5,00 4 . 56678
Ancho del tallo: 10
Cada hoja: 1 caso(s)

Figura 13. Diagrama de tallo y hojas de la variable EFT

Los diagramas de tallos y hojas consisten en una organización gráfica de los números, lo que permite visualizar y tener el conjunto de datos completo, a la vez que observar las características más importantes de la distribución de datos.

Para representar los datos a través de un diagrama de tallo y hojas, cada valor numérico se divide en dos partes. El dígito correspondiente a la posición de las unidades será la "hoja", mientras que los dígitos principales (a la izquierda de la hoja) constituyen el "tallo".

En la figura 13 se observan diez "tallos" o filas (0,0,1,1,2,2,3,3,4,4) que quedan ordenados a la izquierda, cada uno de los cuales tiene una determinada cantidad de "hojas", que correspondería a la frecuencia absoluta de los datos. Antes de esto, se define que hay tres valores "extremos" y menores que 0 (este concepto se estudiará más adelante).

Después de esa primera fila se describe el primer tallo, el 0, que agrupa dos valores, representados por sus hojas (33), que son 3 y 3. Esto significa que dos personas tuvieron un puntaje de 3. El segundo tallo, también 0, agrupa once valores representados por las hojas (55566677888); al añadirles el tallo correspondiente se muestra que los datos originales fueron los equivalentes a las hojas: 5, 5, 5, 6, 6, 6, 7, 7, 8, 8 y 8. El tercer tallo, de 1, contiene 25 hojas (55555666667777777788888999999). Al pegar el tallo con cada hoja podemos saber exactamente qué valores están presentes en los datos (hay cinco casos con puntaje 15, cinco con puntaje 16, ocho con puntaje 17…).

Una ventaja del diagrama de tallo y hojas sobre la distribución de frecuencias en histogramas es que no se pierde la identidad de cada observación. Todos los valores quedan representados en la gráfica.

Las tablas, los histogramas, los diagramas de tallo y hojas y los polígonos de frecuencias describen lo que conocemos como la *distribución de frecuencias* de una variable; esto es, la forma en que las frecuencias se distribuyen a lo largo del rango de la variable. Las formas de la distribución de frecuencias serán importantes para comprender el comportamiento de la variable. Más adelante, veremos que ciertas formas específicas son condición necesaria para que se puedan correr algunas pruebas estadísticas sobre las variables. Por esta razón debemos poder describirlas.

Tipos de distribuciones de frecuencias

Las formas de las distribuciones de frecuencias pueden ser clasificadas de acuerdo a diferentes criterios, y la mayoría de estos provienen de comparaciones que hacemos con una curva de gran importancia en la naturaleza y bastante conocida en la estadística: la curva normal. Esta curva, que tiene una distribución característica en forma de campana, tendrá gran importancia para nosotros más adelante, cuando abordemos los temas de la estadística inferencial. Por ahora, baste decir que la curva normal tiene una distribución de frecuencias unimodal, simétrica y mesocúrtica. En esta parte explicaremos qué significa cada una de estas denominaciones.

Por el número de modas

Observemos de nuevo el polígono de frecuencias de la variable EFT, que volvemos a reproducir en la figura 14a, para facilitar la comparación. Este tipo de polígono se da con bastante asiduidad en la investigación social y educativa. Primero, notemos que existe un área, localizada en una parte central, que acumula la mayoría de los casos de la variable. Este tipo de distribución, que presenta una especie de montaña con una sola cima, o pico, se conoce como *distribución unimodal*; esto es, que presenta una sola moda. La *moda*, como la definiremos en el siguiente capítulo, es el valor más frecuente de una variable.

Aunque la mayoría de las variables en la investigación social y educativa son unimodales, existen algunas que muestran otros comportamientos. Examinemos, por ejemplo, la distribución de frecuencias de la variable "edad", en nuestra muestra de colegios de Bogotá, que se reproduce en la figura 14b.

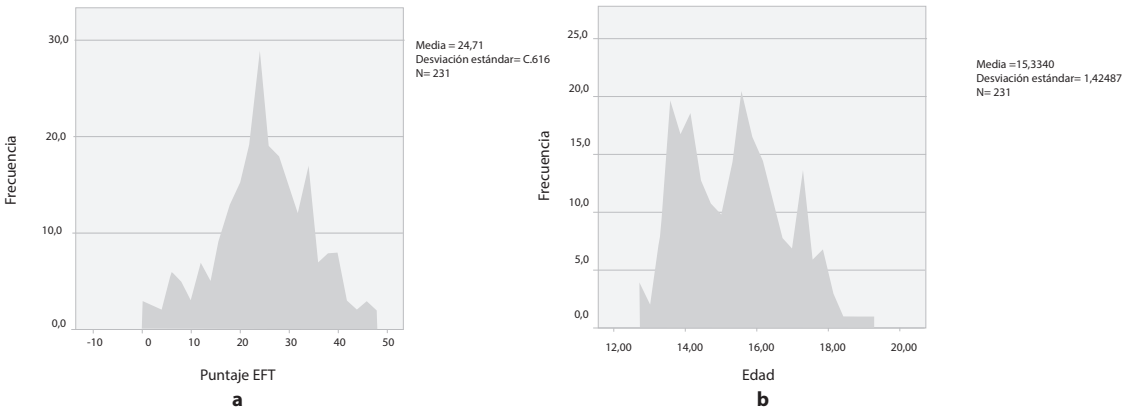


Figura 14. Dos polígonos de frecuencias en una muestra de colegios de Bogotá

Nota: a) polígono de frecuencias de la variable EFT (unimodal); b) polígono de frecuencias de la variable "edad" (bimodal).

Como se observa, la edad parece mostrar no una, sino dos modas diferentes en nuestra muestra. La primera se alcanza antes de los 13 años y, a partir de allí, se nota un descenso importante en el número de casos. La segunda moda se presenta hacia los 16 años y de nuevo, a partir de allí, desciende el número de casos. La presencia de estos dos picos hace que describamos esta como una *distribución bimodal*. Por supuesto, esto es en términos aproximados: que el primero de los picos sea levemente más bajo que el segundo no impide que describamos su comportamiento como prácticamente bimodal.

Puede ser ilustrativo para el lector explicar este comportamiento. Nuestra muestra específica fue obtenida de dos grados diferentes en colegios públicos de Bogotá: los grados 8.= y 10.=. Esta bimodalidad es el resultado directo de esta selección, en la medida en que la edad está estrechamente asociada con el grado que se está cursando. Si segmentamos la base por grado y volvemos a examinar la distribución de frecuencias, encontraremos que las distribuciones cambian. ¿Son ahora unimodales? La respuesta no es rotundamente afirmativa. Veamos los resultados (figura 15).

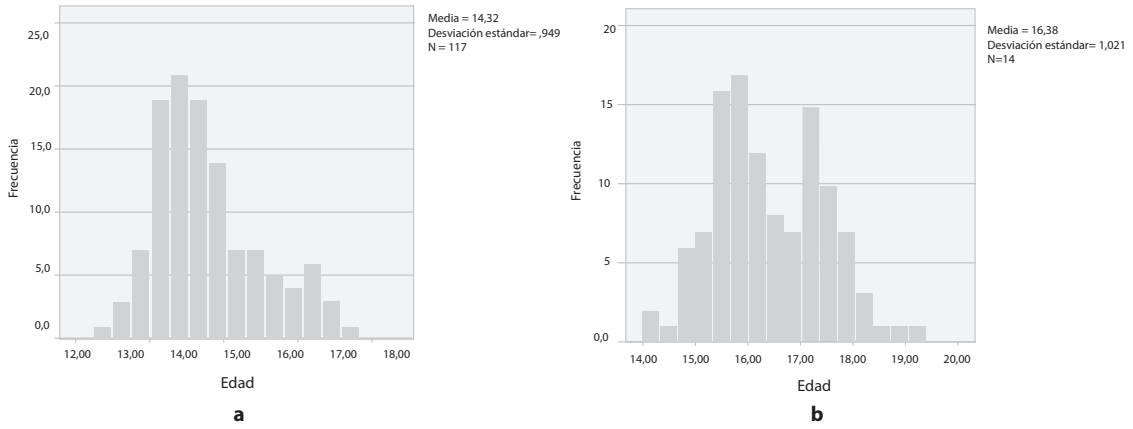


Figura 15. Histogramas de la edad para la muestra

Nota: a) grado 8.°; b) grado 10.°.

Como se observa, la distribución de frecuencia de la edad para los estudiantes de grado 8.°= muestra una tendencia claramente unimodal, entre los 13 y los 14 años, edad esperada para el grado, si bien se alcanza a presentar una larga cola derecha conformada por estudiantes que, estando matriculados en el grado, muestran edades mucho más altas que las esperadas. En el caso de la gráfica correspondiente al grado 10.° se vuelve a presentar una tendencia aproximadamente bimodal. Una primera moda se da alrededor de la edad esperada para el grado, entre los 15 y los 16 años, y una segunda, entre los 17 y los 18 años. Pareciera que, en este grado, se empieza a notar una proporción importante de estudiantes con "extraedad educativa" que podría diferenciarse de los estudiantes regulares en otros muchos indicadores.[3]

3 El concepto de *extraedad educativa* hace referencia a los estudiantes que, por diferentes razones, presentan una edad mucho mayor que la correspondiente, o esperada, para el grado que se encuentran cursando. Con frecuencia revela efectos acumulados de otros fenómenos: deserción parcial, mortalidad académica, repitencia, entre otros.

Además de las distribuciones unimodales y bimodales, se habla de distribuciones multimodales para referirse, en general, a los casos de más de dos modas en una distribución.

Todavía es posible describir otra forma de distribución en términos del número de modas presentes. En efecto, existen algunas distribuciones en las que no se puede reconocer ninguna moda a lo largo del rango de la variable. Esta distribución, en la que todos los valores presentan, aproximadamente, la misma frecuencia, se conoce como *distribución rectangular.*

Por la simetría

Consideremos ahora otro criterio para la caracterización de distribuciones, esta vez relacionado con la *simetría* de la distribución. Si se observa la gráfica del puntaje EFT de la figura 14a, es fácil constatar que la distribución muestra la moda en el centro y, a izquierda y derecha, se encuentra un número similar de casos. Esto es lo que conocemos como *distribución simétrica*. No todas las distribuciones lo son. Observe, por ejemplo, los histogramas correspondientes a las variables "número de colegios en los que se ha estudiado" (figura 16a) y "número de personas que viven con el estudiante" (figura 16b).

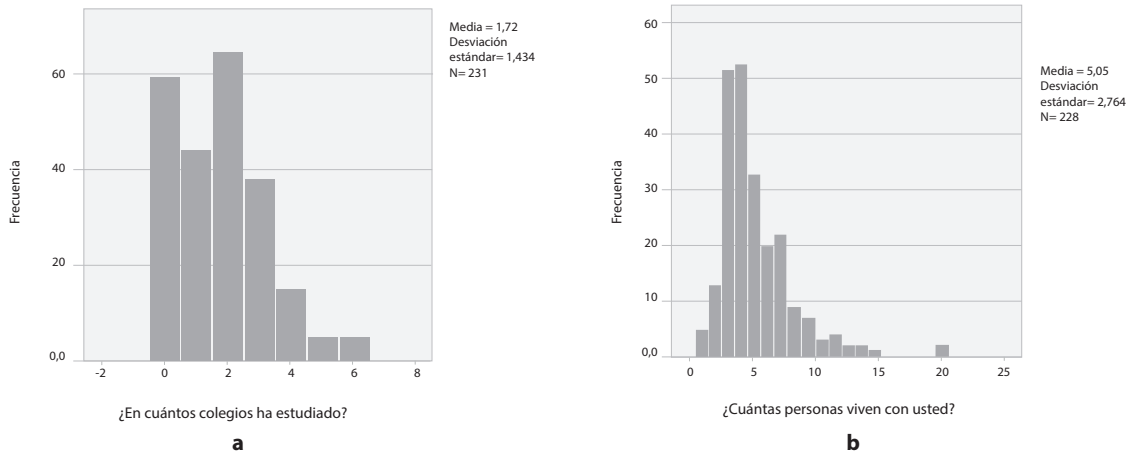


Figura 16. Dos distribuciones positivamente asimétricas

Nota: a) ¿en cuántos colegios ha estudiado?; b) ¿cuántas personas viven con usted?

Tal y como se observa en la figura 16a, muchos estudiantes dentro de la muestra han permanecido en un solo colegio durante toda su historia escolar. Existe también un número importante de personas que ha cambiado una e incluso dos veces, y a partir de allí la curva de la distribución empieza a bajar rápidamente hasta llegar a un máximo de seis.

La figura 16b muestra el mismo comportamiento, pero con mayor dispersión de valores. Al responder a la pregunta ¿cuántas personas viven con usted?, la mayoría de los estudiantes dicen que dos o tres personas y, a partir de allí, la curva desciende formando una larga cola a la derecha que llega hasta el número 20. En estos casos, en los que tenemos esa marcada asimetría con largas colas a la derecha, diremos que estos son ejemplos de distribución *asimétrica hacia la derecha*, o *positivamente asimétrica*.

Es relativamente fácil encontrar, en la investigación educativa y social, distribuciones positivamente asimétricas. Básicamente se trata de hallar una característica que tenga un límite inferior: el número de personas que viven con el estudiante, el número de libros leídos en un año o el número de hijos son ejemplos típicos (no se pueden tener menos de cero hijos). Esto se conoce como *efecto piso*.

Existen también *distribuciones asimétricas hacia la izquierda*, o con *asimetría negativa*. El caso, por ejemplo, de la puntuación de un examen que resultó especialmente fácil para una muestra de estudiantes, podría mostrar una distribución asimétrica hacia la izquierda y un *efecto techo*.

Los histogramas contenidos en la figura 17 son dos ejemplos de distribuciones asimétricas a la izquierda. La información proviene de la aplicación de un instrumento sobre creencias acerca del aprendizaje en un grupo de 1017 estudiantes de carreras de educación. Los resultados se presentan en una escala numérica de 1 a 5, donde "1" representa total desacuerdo y "5" representa total acuerdo.

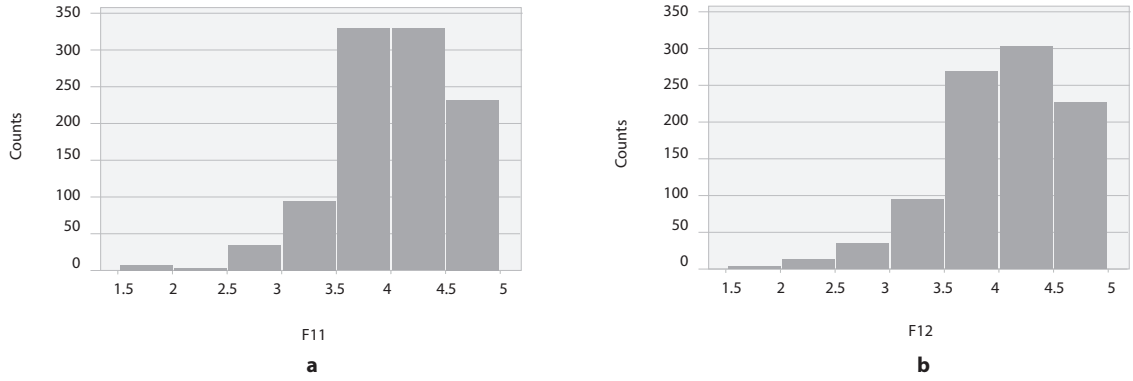


Figura 17. Dos distribuciones con asimetría negativa

Nota: a) aprender es construir conocimiento; b) aprender es poder usar el conocimiento.

El histograma de la figura 17a (F11) representa la creencia de que el aprendizaje es un proceso de construcción; el de la figura 17b (F12), la idea de que conocer algo es poder utilizarlo. Como se observa, los resultados plasmados en las gráficas muestran que los estudiantes de carreras de educación creen, mayoritariamente, en esas dos ideas.

Por el grosor de las colas

Por último, es también posible clasificar la distribución de frecuencias de acuerdo con lo relativamente gruesas o delgadas que sean sus colas. Este criterio se conoce como *curtosis*. El término proviene de la palabra griega *kyrtos* que significa 'curva'. De acuerdo con el criterio de la curtosis, es posible diferenciar las distribuciones de colas gruesas, también llamadas *leptocúrticas*, las curvas con un grosor intermedio (como la curva normal), llamadas *mesocúrticas*, de las curvas con colas delgadas, llamadas *platicúrticas* (figura 18).

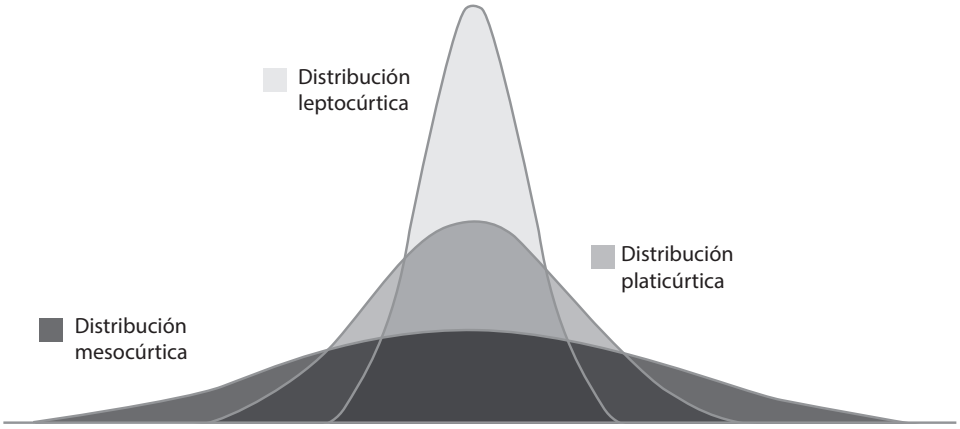


Figura 18. Distribuciones simétricas con diferentes niveles de curtosis

Habitualmente, las distribuciones leptocúrticas son más empinadas que la distribución normal, mientras que las platicúrticas son más achatadas (como platos invertidos). De cualquier manera, el criterio más importante es la forma de las colas.

Criterios numéricos para el examen de distribuciones

Existen criterios numéricos para evaluar el nivel de asimetría o curtosis de una distribución de frecuencias. En los diferentes programas que utilizamos, es posible solicitar que el programa calcule los valores de la asimetría (*skewness*) o la curtosis (*kurtosis*) en el mismo menú en el que se calcula la tabla de frecuencias.

Para la interpretación de los valores de la asimetría y la curtosis arrojados, debe tenerse en cuenta lo siguiente:

- Asimetría. La distribución normal es simétrica y el valor de su asimetría es 0, si bien con frecuencia se asumen simétricos los valores absolutos de la asimetría menores que 0,5. Si una distribución tiene asimetría positiva, tiene una cola derecha larga. Como regla general un valor de la asimetría mayor, en valor absoluto, que el doble de su error estándar se asume que indica una desviación en la simetría.
- Curtosis. La distribución normal tiene una curtosis de 0. Como en el caso de la asimetría, se aceptan como mesocúrticos valores absolutos de curtosis menores que 0,5. Una curtosis positiva indica una curva leptocúrtica; una negativa indica una curva platicúrtica. Como en el caso de la asimetría, un valor de la curtosis mayor, en términos absolutos, que el doble de su error estándar indica una desviación de la curtosis de la curva normal.

En las gráficas de la figura 19, obtenidas de diferentes investigaciones, se muestran valores diferentes de asimetría y curtosis que pretenden ser ilustrativos. Las gráficas de la izquierda (figuras 19a, 19c y 19f) muestran curvas con asimetría negativa; las centrales (figuras 19b y 19d), curvas simétricas, y las de la derecha (figuras 19c y 19g), curvas con asimetría positiva. En otra dimensión, las

gráficas de la fila superior (figuras 19a y 19b) muestran valores negativos de la curtosis, mientras que las de la fila inferior (figuras 19f y 19g) muestran curtosis positivas.

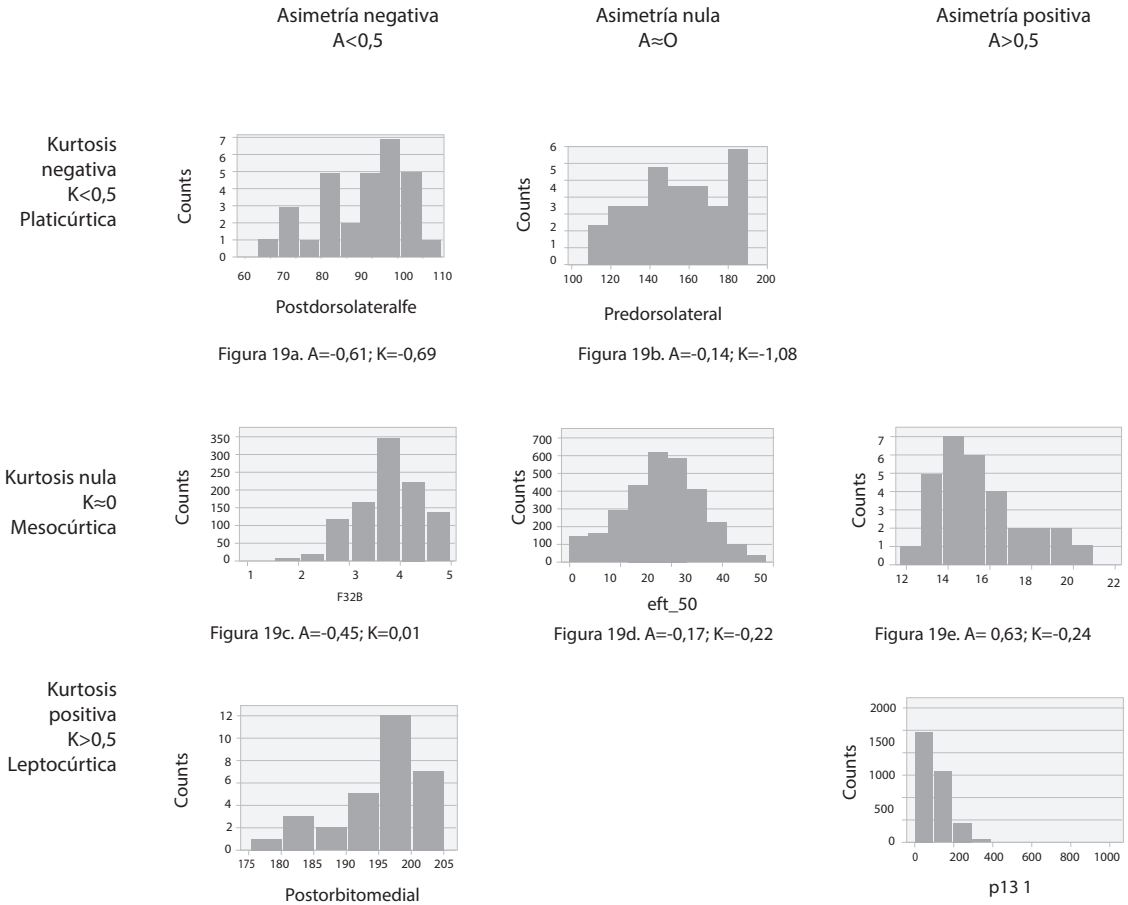


Figura 19a. A=-0,61; K=-0,69

Figura 19b. A=-0,14; K=-1,08

Figura 19c. A=-0,45; K=0,01

Figura 19d. A=-0,17; K=-0,22

Figura 19e. A= 0,63; K=-0,24

Figura 19. Valores de asimetría (A) y curtosis (K) en diferentes distribuciones

Percentiles

El concepto

Los *percentiles* son medidas de posición relativa de un puntaje frente al resto de puntajes en la muestra. Se utilizan con frecuencia, en el ámbito de la educación, para comparar el rendimiento de un individuo frente al de su grupo de referencia. Debido a que sirven básicamente para comparar, requieren variables con niveles de medida, al menos, ordinales.

Supongamos, por ejemplo, que un estudiante obtiene un puntaje de 123 en un examen. ¿Esto es alto? ¿Bajo? Por sí mismo, esto es imposible de interpretar si ni siquiera conocemos la escala. Supongamos ahora que la escala en la que se califica el examen es, digamos de 0 a 150. Esto nos indica algo más: nos podría señalar que el puntaje de 123 podría ser alto, considerando que se sitúa más cerca del extremo superior de la escala.

Sin embargo, los datos del puntaje y de la escala no son suficientes. Para interpretar con mayor claridad el puntaje necesitamos saber cómo fueron los puntajes del resto de personas que tomaron

el examen. Si, por ejemplo, ahora supiéramos que un puntaje de 123 es superior al 98 % de los puntajes de la clase, diríamos con mayor certeza que es un excelente puntaje. Si, por el contrario, supiéramos que este mismo puntaje es inferior al 98 % de los puntajes de la clase tendríamos que concluir que, en términos comparativos, no fue un buen puntaje. Para esto son los percentiles: para comparar puntajes concretos con el rendimiento global.

Ahora, comparar un puntaje concreto con los puntajes obtenidos por las otras personas que tomaron el examen nos permite interpretarlo en términos relativos, no absolutos. Por ejemplo, el que un sujeto haya obtenido, en un examen de conocimientos, un puntaje que supera al 98 % de su clase, no quiere decir que sepa mucho, solo que sabe más que la mayoría. Y a la inversa, un sujeto puede estar en el percentil 5 (el 95 % de la clase tiene un puntaje mayor que él) y, aun así, saber lo suficiente para aprobar la materia. Por esta razón los percentiles pueden ser, aunque útiles, muy discutibles si los utilizamos como dato único para establecer calificaciones.[4]

Para lo que nos ocupa en este contexto, necesitamos comprender el concepto de punto percentil y poder dividir una muestra en algunos grupos percentiles de uso frecuente.

Iniciemos con las definiciones básicas. Un *punto percentil* es un valor, sobre la escala de medición, debajo del cual se encuentra un porcentaje dado de los datos. Por ejemplo, el percentil 40 es el punto de la escala por debajo del cual se encuentra el 40 % de los datos. Algunos puntos percentiles permiten la definición de grupos percentiles. Un *grupo percentil* es el conformado por todos los sujetos contenidos entre dos puntos percentiles sucesivos.

Existen algunos puntos percentiles notables, por lo frecuente de su uso, lo que les ha valido tener nombres específicos: deciles, terciles, cuartiles y quintiles son los más mencionados.

- Los deciles son los puntos percentiles en el rango de las decenas (10, 20, 30, etc.) y se simbolizan como D1, D2, etc. También se utiliza el término decil para designar cada uno de los grupos percentiles contenidos entre dos puntos sucesivos, que contendrá el 10 % de los datos. En este caso, decimos que el primer decil contiene el primer 10 % de los datos; el segundo decil, el siguiente 10 %, y así sucesivamente.
- Los terciles son los puntos percentiles que cortan la muestra en tres grupos iguales, que contienen el 33,33 % cada uno. De igual forma, designan cada uno de los tres grupos percentiles correspondientes.
- Los cuartiles son los puntos percentiles que dividen al grupo en cuatro partes con idéntico número de sujetos (25 %). Se simbolizan con Q1, Q2, Q3 y Q4. El cuartil Q2 (segundo cuartil) corresponde al percentil 50 (más adelante se le llamará *mediana*, que se utiliza como medida de tendencia central); divide al grupo en dos partes iguales. Los cuartiles se calculan a veces como puntos de referencia y para hacer determinadas representaciones gráficas, tales como los diagramas de cajas y bigotes, que examinaremos más adelante. Como en los casos anteriores, en este también utilizaremos palabra "cuartil" para designar, no solo el punto de corte, sino cada uno de los grupos de datos generados por estos puntos.

4 Puede ser útil que el lector recuerde la información sobre los diferentes tipos de evaluación y, específicamente, sobre las diferencias entre una evaluación *con referencia a la norma* (en la que se comparan los resultados de las personas, entre sí) y una evaluación hecha *con referencia a criterio* (en la que se compara en cada ejecución un criterio de lo que se considera adecuado).

- Los quintiles son la extensión del concepto anterior para dividir el grupo en cinco partes iguales, que contenga cada una el 20 % de los datos.
- En general, se habla de *n-tiles*, para designar cada uno de los *n* grupos percentiles formados por dos puntos *n-tiles* sucesivos.

Por último, es importante anotar que el percentil no trata de una puntuación propiamente dicha, puesto que no está referido a la variable que se ha medido; no hay una unidad. Entre dos percentiles contiguos no hay la misma distancia en aquello que estamos midiendo. Así, si un estudiante, en un examen, está en el percentil 80, no podemos decir que sabe el doble del que esté en el percentil 40; solo que esta puntuación tiene, por debajo, el doble del número de puntuaciones.

Obtener los puntos percentiles y dividir la muestra en n grupos iguales

Para conocer los puntos percentiles de una determinada variable utilizando el SPSS, el camino más sencillo es a través del procedimiento de frecuencias, que ya hemos estudiado (/Analizar/Estadísticos descriptivos/Frecuencias…). En este menú, en el botón "Estadísticos", es posible solicitar los cuartiles, puntos de corte para un número cualquiera de grupos iguales, o cualquier percentil determinado.

Cuando se solicitan, por este medio, los cuartiles de la variable que hemos venido utilizando como ejemplo (puntaje EFT), el programa, arroja la tabla 7.

Tabla 7. Salida del spss al solicitar cuartiles en el menú "frecuencias"

Estadísticos		
Puntaje EFT		
N	Válido	231
	Perdidos	0
Percentiles	25	19,00
	50	24,00
	75	31,00

Para dividir la muestra en grupos *n*-tiles, el SPSS dispone del procedimiento "Asignar rangos a casos", disponible en la dirección "/Transformar/Asignar rangos a casos…". En este menú, después de seleccionar la variable correspondiente, se activa el botón "Tipos de rango", que permite multitud de posibilidades, entre las cuales existe "Ntiles", en la que puede designarse el número de grupos iguales a voluntad. Este procedimiento crea una nueva variable en la que, a cada caso, se le asigna su valor *n*-til correspondiente. Si, por ejemplo, solicitamos n-tiles "4", la nueva variable tendrá solo cuatro valores y a cada caso se le asignará el valor correspondiente a su cuartil. Cuando se pide la tabla de frecuencias de esta nueva variable, el SPSS arroja la salida de la tabla 8.

Tabla 8. Frecuencias de la variable NEFT50, construida con los grupos cuartiles de la variable EFT

NEFT50 percentile group of EFT50					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válido	1	60	26,0	26,0	26,0
	2	57	24,7	24,7	50,6
	3	57	24,7	24,7	75,3
	4	57	24,7	24,7	100,0
	Total	231	100,0	100,0	

Observe que, en este caso particular, no fue posible construir grupos que contuvieran exactamente el 25 %. El programa arroja la mejor aproximación posible.

Representaciones gráficas

Las gráficas más usadas para la representación de los puntos percentiles son los llamados *diagramas de cajas* (*box plots*). Estos diagramas logran agrupar los datos en cuatro secciones que corresponden a los cuatro cuartiles.

Para el caso de la variable que hemos venido trabajando, y que corresponde a los puntajes obtenidos en una muestra de 231 personas en la prueba EFT, el diagrama de cajas (también conocido como diagrama de cajas y bigotes), tal y como lo hace el SPSS, se muestra en la figura 20.

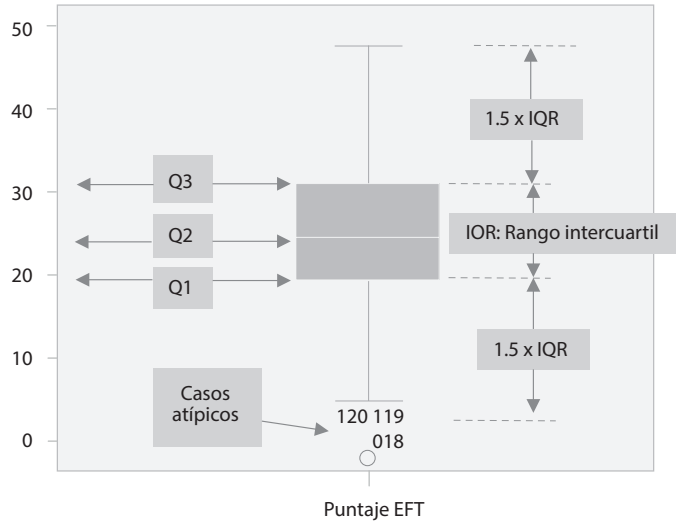


Figura 20. Diagrama de cajas de la variable EFT

Para la interpretación de este diagrama, se supone que la línea negra gruesa dentro de la caja representa el segundo cuartil (Q2), o el punto en donde se divide el primero y el segundo 50 % de los datos. Este valor pasará a ser definido más adelante como la mediana, y en nuestro caso corresponde al punto 24.

Los extremos de la caja representan los cuartiles 1 y 3 (Q1 y Q3). En nuestro caso, estos valores son 19 y 31, respectivamente (ver tabla 7). Esto significa que, dentro de la caja están contenidos el 50 % de los datos más "centrales": los que se encuentran por encima del primer 25 % y por debajo del último 25 %. La longitud de la caja se conoce como *rango intercuartil* (o *interquartil range*, IQR), y es una medida de dispersión de los datos relativamente usada. En nuestro caso, el rango intercuartil de esta variable es de:

$$\text{Rango intercuartil} = Q3\text{-}Q1 = 31\text{-}19 = 12$$

En los extremos de la caja aparecen unas barras en forma de "T", conocidas como los "bigotes". Tienen una extensión de 1,5 veces el rango intercuartil. Estas líneas intentan expresar gráficamente la localización de los datos contenidos en el primer rango cuartil (Q1), desde el valor más bajo, y en el último (Q4) hasta el valor más alto.

Los programas de procesamiento de datos representan algunos datos puntuales, conocidos como *casos extremos*, anotando, en la gráfica de cajas, el número de caso correspondiente para su identificación en la base. Un valor es considerado extremo cuando se sitúa por fuera de los bigotes. Existen dos tipos de casos extremos: el primero, visible en nuestra gráfica, se representa con pequeños círculos, es conocido como *caso atípico* y son valores que se encuentran por fuera de los bigotes, pero a una distancia de la mediana menor que tres rangos intercuartiles. El segundo es llamado *caso atípico extremo* y se representa por asteriscos; es aquel que se sitúa a más de tres rangos intercuartiles de la mediana.

Uso de las tablas de frecuencia y sus gráficas en publicaciones científicas

Los investigadores sociales y educativos elaboran y usan tablas de frecuencias y las expresan en gráficas e histogramas. Se utiliza, incluso, como procedimiento para ayudar a depurar la base de datos, eliminando aquellos claramente inconsistentes, o fuera de rango, que pudieran estar allí por errores de codificación o digitación. Estas tablas y gráficas se usan también para ayudar al analista a comprender e ir incorporando una visión general de los datos.

A pesar de que siempre se hacen estas tablas y gráficas, muy rara vez pasan a publicarse en un artículo científico, básicamente porque se utilizan como pasos previos para la preparación de estadísticos más elaborados. Muchas pruebas estadísticas requieren de ciertos supuestos sobre las formas de la distribución de frecuencias las variables involucradas, y existen técnicas, que veremos más adelante, para examinar y asegurar el cumplimiento de estos supuestos.

Por otro lado, es fácil reconocer que, cuando en un artículo se describen las características de la muestra efectiva, se han examinado frecuencias y estas se han expresado, usualmente, en textos. Cuando las variables son categóricas, nominales u ordinales, la frecuencia es casi la única forma de describirlas. Observe el siguiente fragmento de un artículo real:

> *El estado civil soltero (96 %) predomina en los estudiantes, y, en menor porcentaje se encuentran estudiantes casados (2,4 %), en unión libre (1 %) y separados (,6 %). Respecto al estrato socioeconómico, tal y como lo define el Departamento Nacional de Estadística*

de Colombia (DANE), la mayoría de los estudiantes residen actualmente en viviendas que corresponden a un estrato socioeconómico identificado como 2 —o medio bajo (35 %)—, seguido del estrato 3 —bajo (23,8 %)—, el estrato 4 —medio (18 %)—, y el estrato 5 —medio alto (1,4 %)—. (Caballero et al., 2015, p. 5)

Existen también algunos casos en los que la finalidad de la investigación es, básicamente, de tipo descriptivo. En estos casos podremos encontrar tablas de frecuencias simples, o agrupadas, presentadas como resultados principales. Ese es el caso de estudios bibliométricos, como el del ejemplo que sigue:

Con respecto al país en el que están ubicadas las instituciones representadas por los diferentes autores, en Colombia está el 36,9 %. Otros países que muestran buenos niveles de participación en la revista son España (18,4 %), Chile (11,5 %), México (10,6 %), Argentina (6,0 %) y Brasil (6,0 %). Entre estos seis países se ubica casi el 90 % de la participación. Las instituciones ubicadas en otros países muestran, cada una, menos del 2 % de participación.

País	N.º	(%)
Alemania	1	0,5
Argentina	13	6,0
Australia	2	0,9
Brasil	13	6,0
Canadá	1	0,5
Chile	25	11,5
Colombia	80	36,9
Ecuador	2	0,9
España	40	18,4
Estados Unidos	5	2,3
Francia	1	0,5
Grecia	1	0,5
Inglaterra	1	0,5
Israel	1	0,5
México	23	10,6
Perú	1	0,5
Portugal	3	1,4
Reino Unido	2	0,9
Uruguay	1	0,5
Venezuela	1	0,5
Total	217	100,0

(*Hederich-Martínez y Roa-Casas, 2019, p. 216).*

Capítulo 3

Medidas de tendencia central, dispersión y puntuaciones Z

Presentación

Las medidas estadísticas univariadas son de tres tipos: 1) *las medidas de tendencia central*, usadas para resumir, en un solo número, un grupo de datos; 2) *las medidas de dispersión*, utilizadas para describir la variación de los datos con respecto a las medidas de tendencia central; y 3) *las medidas de posición*, utilizadas para describir la posición relativa de un individuo o grupo de individuos en distribución del total de los datos.

En este capítulo explicaremos las medidas de tendencia central y las de dispersión, junto con una transformación que sintetiza las dos en un solo número: las puntuaciones Z. Las medidas de posición ya fueron abordadas en el capítulo anterior cuando hablamos de percentiles.

Medidas de tendencia central

Media aritmética

En general, la medida más común y popular para describir un grupo de observaciones numéricas es su promedio, llamado en estadística *la media*. Definida brevemente, la media de un grupo de observaciones numéricas es la suma de todos los valores dividido por el número de observaciones.

$$M = \sum_1^n x_i / n$$

En formato APA, la media muestral se representa mediante el uso de las letras *M* o *X* (*X barra*); la media poblacional, por su parte, se representa mediante el uso de la letra griega μ (se lee "miu").

A muchas personas les resulta útil imaginar la media como el punto de equilibrio de la distribución de observaciones. Imagine, por un momento, pilas de cubos que representan el histograma de la distribución de la variable, puesta sobre una tabla sin peso. En esta situación, la media quedaría ubicada en el punto en donde la tabla se encuentra en equilibrio.

Para el caso del ejemplo en el que hemos venido trabajando, del puntaje EFT, la media, como punto de equilibrio, se vería como se muestra en la figura 21.

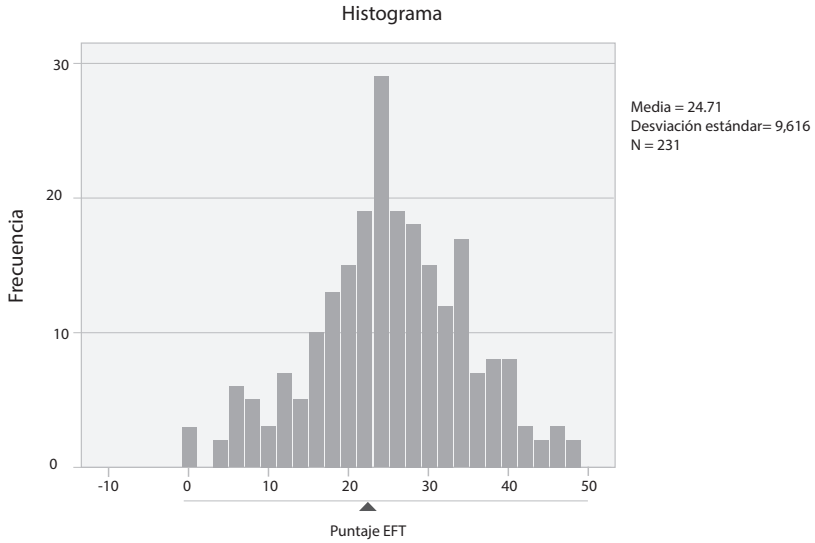


Figura 21. La media como punto de equilibrio

Observe que el valor de la media no tiene que estar presente entre los valores de la muestra. En nuestro caso, aunque la variable es discreta, la media (M=24,71) presenta decimales. Ninguna persona obtuvo un puntaje de 24,71 en esta prueba.

Existen varias consideraciones de importancia al usar medias:

- Utilice medias solo en los casos de variables propiamente numéricas, sean estas de intervalo, de razón, discretas o continuas. El uso de medias en variables ordinales, aunque bastante generalizado, no está muy recomendado: para ello, se dispone de otros tipos de medida de tendencia central, de las que hablaremos más adelante. El uso de medias en variables nominales es claramente absurdo; en este tipo de variables los números son solo códigos que no representan cantidades y, por tanto, no pueden siquiera ser sumados.
- Es muy importante anotar que la media se ve muy afectada por las puntuaciones extremas; no así otras medidas de tendencia central que veremos a continuación, como la mediana y la moda. Por ejemplo, en un curso de tercero de primaria donde todos los niños tienen una edad que oscila entre 8 y 9 años y uno solo tiene 12 años, esta última puntuación extrema afecta la media del curso, mostrando una media engañosamente más alta.

Mediana

En el caso de tener entre los datos un valor muy grande o muy pequeño con respecto a los demás, se puede obtener una media no muy representativa del conjunto de datos. En este tipo de situaciones, la tendencia central de este tipo de datos es mejor descrita por otra medida de tendencia central denominada la mediana.

La *mediana* es el punto medio de los datos al ordenarlos por su valor, ya sea de forma ascendente o descendente; esto es, la mediana es el dato que presenta el mismo número de datos por encima o por debajo. Se considera la mediana como una medida de orden y para obtenerla se requiere que, como mínimo, la variable tenga un nivel de medida ordinal.

Un punto adicional para el cálculo de la mediana. Cuando tenemos un número impar de datos diferentes, obtener la mediana será sencillo, ya que esta coincide con el dato que está en el centro. Pero ¿cómo debemos proceder cuando tenemos un número par de datos? La tabla 9 aclara el proceso.

Tabla 9. Procedimiento para calcular la mediana con un número impar o par de datos

Datos impares	Datos pares
20 22 26 33 34 ⇨ 39 88	20 22 26 29 30 ⇦ 96
En este caso, la mediana es 33, ya que es el cuarto dato entre siete datos en total. Quedan tres datos por encima, y por debajo de la mediana.	En este caso no hay un dato a elegir como mediana, ya que la condición es que la cantidad de datos por encima y debajo de la mediana sean iguales. La solución en este caso es tomar los dos datos medios y sacar el promedio entre estos dos $Mdn = \frac{Dato3+Dato4}{2} = \frac{26+29}{2} = 27{,}5$

Esta medida de tendencia central resulta especialmente útil cuando tenemos distribuciones asimétricas, en las que existen algunos pocos datos atípicamente altos o bajos. En distribuciones simétricas, la media y la mediana coinciden y debería emplearse siempre la media.

Un ejemplo típico de uso de la mediana como medida de tendencia central en la investigación social se da en datos relacionados con el nivel socioeconómico, y específicamente con el nivel de ingresos de una familia. Como se sabe, en este tipo de variables se tienden a presentar algunos individuos aislados que presentan muy altos ingresos, por lo cual la media sobreestima el ingreso obtenido por las familias.

Otro de los usos frecuentes de la mediana por encima de la media se da en datos de investigaciones psicológicas en las que se mide el tiempo de latencia de respuesta del individuo frente a una tarea. En este caso ocurre, con alguna frecuencia, que unos pocos individuos en algún momento se dis-

traigan, por lo que la medición de su tiempo de respuesta muestra valores atípicamente altos. En este caso, es preferible utilizar la mediana como medida de tendencia central.

Para expresar el valor de la mediana de un conjunto de datos se utiliza el código *Mdn*.

Moda

La *moda* es el valor que se presenta con mayor frecuencia en una distribución. Esto la hace la medida de tendencia central más fácil de determinar, ya que no requiere de ningún cálculo; en general, una simple inspección visual de la distribución será suficiente. En otro momento hemos hablado de distribuciones de frecuencias aproximadamente bimodales, o con dos modas; esto podría extenderse conceptualmente a las distribuciones multimodales, o con múltiples modas.

Como punto a favor, la moda es la única medida de tendencia central que puede ser utilizada en variables con niveles de medida estrictamente nominal. Como la mediana, la moda no se ve influida por los valores muy altos o muy bajos; de hecho, en variables estrictamente nominales el concepto de "alto" o "bajo" no puede ser determinado. Piénsese, por ejemplo en variables como "nacionalidad" o "etnia".

Como punto en contra, ya hemos mencionado que la moda no siempre existe en todos los conjuntos de datos o, dicho de otra manera, pueden existir múltiples modas en el conjunto. Por esta razón, la moda no se utiliza en la misma frecuencia que la media y la mediana.

Medidas de dispersión

Las medidas de tendencia central que hemos expuesto en la sección anterior son muy útiles a la hora de dar cuenta, con un solo indicador, del estado de un conjunto de datos. En efecto, la media, la mediana y, en menor medida, la moda tienen la capacidad de resumir todo el conjunto de datos a un solo valor, lo cual resultará muy útil en términos descriptivos.

Sin embargo, la tendencia central de una distribución es, aunque importante y con gran valor descriptivo, claramente insuficiente a la hora de dar cuenta del conjunto de datos. Para dar una idea más precisa de la distribución debemos dar información acerca del nivel de dispersión de los datos.

Examine las cuatro distribuciones de la variable "edad" medida en años cumplidos en la figura 22. En todos los casos, el conjunto de datos está conformado por 20 casos (n=20) y la media, para los cuatro casos, es la misma: 19,50 años.

A simple vista, es evidente que todas las distribuciones son claramente diferentes. En la figura 22a, tenemos una distribución simétrica, alrededor de un área central en donde estarían media, mediana y moda, que serían iguales a 19,50. Los valores más alejados van descendiendo en frecuencia de forma relativamente suave.

La figura 22b, también simétrica, muestra una tendencia bimodal: las dos áreas que muestran tendencias modales están un poco distantes entre sí: la primera, alrededor de 16 años, y la segunda, alrededor de 23 años. Intuitivamente, podemos constatar que los datos en esta segunda distribución están más dispersos que en la primera.

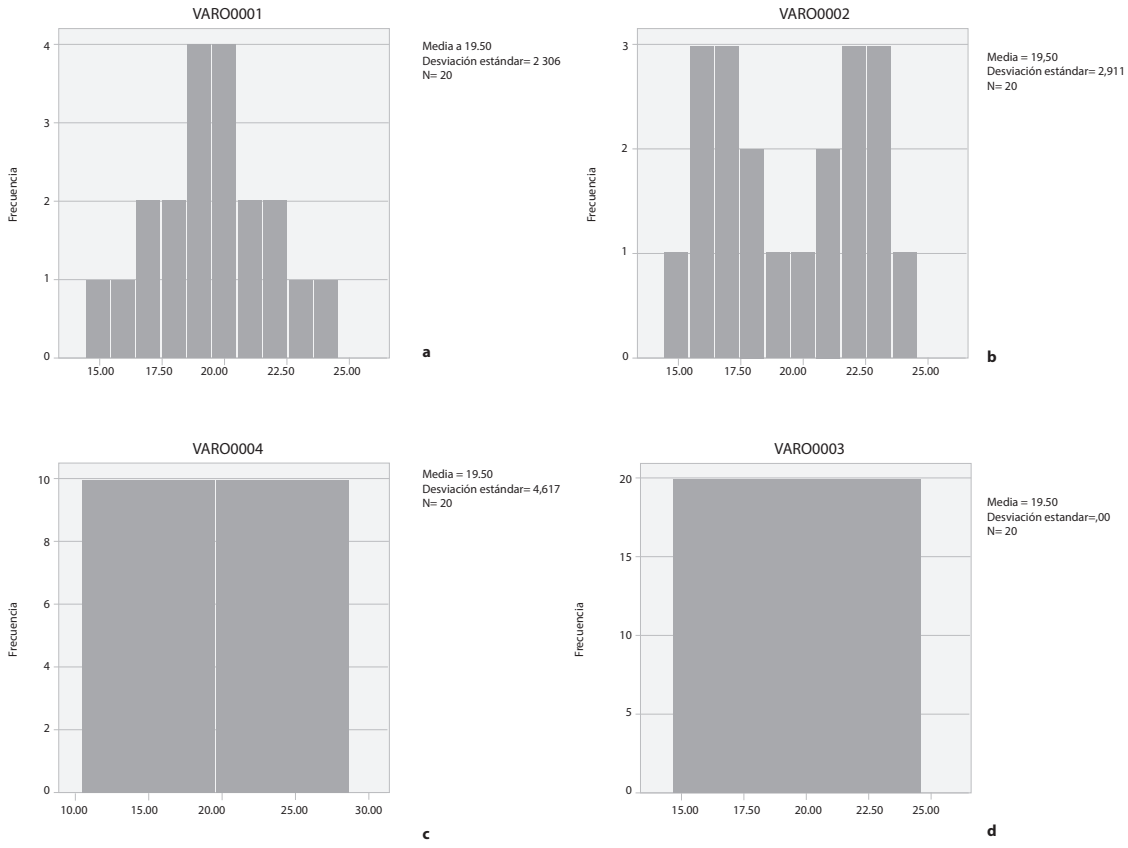


Figura 22. Cuatro distribuciones de la variable "edad"

Nota: a) distribución simétrica unimodal, b) distribución simétrica bimodal, c) dos valores extremos y d) un valor constante.

La figura 22c es un caso más extremo que el anterior. Ahora la mitad de los casos tiene 15 años y la otra mitad tiene 24 años. Es una distribución rectangular con una muy alta dispersión de los datos, aunque la media, al igual que la mediana, sigue siendo 19,50. La distribución de la figura 22d es el otro extremo, pero en el sentido contrario. En este caso, la variable es totalmente homogénea y para cada caso el valor de la edad es igual a 19,50. Por supuesto, la media, mediana y moda son iguales a 19,50, si bien no hay dispersión alguna en el conjunto de datos.

En lo que sigue, se intentará mostrar, de forma comprensiva, diferentes indicadores del nivel de dispersión de los datos que podrían ser útiles a la hora de describir esta característica en el conjunto de datos.

Rango

Este es el primero y más sencillo de todos los indicadores de dispersión en un conjunto de datos. El *rango* es la distancia presente entre el mayor valor de la variable, o su valor máximo, y el menor valor, o mínimo.

Aunque el rango, como medida de dispersión, tiene un significado muy claro, se utiliza muy poco por su marcada inestabilidad. Una sola puntuación extrema, en una u otra dirección, puede modificar de forma muy sustancial el rango de una variable. Por otro lado, si dos variables muestran

distribuciones muy diferentes dentro de los mismos límites, el rango no podrá mostrar esa dispersión "interna" de la variable.

Si ponemos a prueba esta primera medida de dispersión en nuestras cuatro distribuciones ficticias, aunque encontraremos algunas diferencias, será evidente que el rango no es una medida de dispersión muy precisa. En la primera, la segunda y la tercera variable, el rango es el mismo: R = 24 años–15 años = 9 años. Eso sí, en la última variable, con la mínima dispersión, tenemos también el mínimo rango: 0 años.

Rango intercuartil

Ya antes hemos introducido el concepto de rango intercuartil como medida de dispersión de datos ordinales cuando hablamos de diagramas de cajas. Sea este el momento de introducirlo formalmente. Como se recordará, el *rango intercuartil* (RIC o IQR, en inglés) es la distancia entre el valor percentil 25, o primer cuartil (Q1), y el valor percentil 75, o tercer cuartil (Q3). En otras palabras, el rango intercuartil corresponde al rango del 50 % intermedio de los datos.

Puede ser interesante observar los diagramas de cajas y bigotes de las cuatro edades ficticias que venimos examinando, presentados en la figura 23. Como se recordará, el rango intercuartil queda representado en estos diagramas por la longitud de la caja central. Tal y como se observa, este tipo de diagramas tiene un enorme potencial para expresar, al tiempo que la tendencia central, la dispersión de los datos.

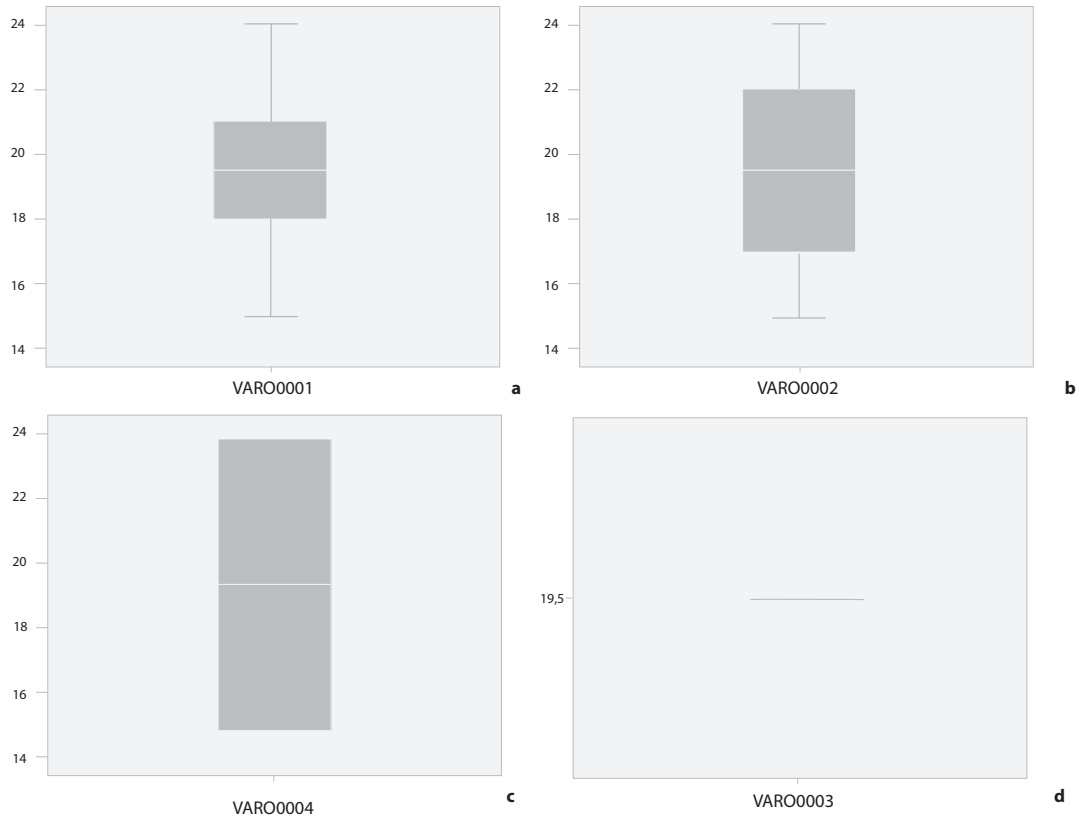


Figura 23. Diagramas de cajas y bigotes de cuatro variables de "edad"

Nota: a) distribución simétrica unimodal; b) simétrica bimodal; c) dos valores extremos; y d) un valor constante.

La primera de las variables que hemos examinado, unimodal alrededor de un punto central, muestra un rango intercuartil de 3. En la segunda variable teníamos un comportamiento bimodal, y el rango intercuartil es ahora de 5. Obsérvense las visibles diferencias entre las longitudes de caja. En la tercera, el rango y el rango intercuartil coinciden en un valor de 9. En la cuarta, estos dos indicadores igualmente coinciden, pero ahora en un valor de 0.

Como se observa, el rango intercuartil es un excelente descriptivo del nivel de dispersión de un conjunto de datos; mucho más preciso y elocuente que el rango de la variable. Cuando se trata de describir la dispersión del conjunto de datos, el rango intercuartil se usa a menudo con datos de escalas preferentemente ordinales. Sin embargo, el rango intercuartil tiene, en general, muy poco uso en estadística, puesto que no se ha utilizado en las pruebas de la inferencia estadística, que examinaremos al final de este libro. Por esta razón, no dedicaremos más atención a este indicador de dispersión para dedicarnos a los que utilizaremos en las pruebas de la estadística inferencial.

Desviación estándar y varianza

Los estadísticos más utilizados para la descripción de la dispersión de una variable y con un amplio uso en la inferencia estadística son la varianza y la desviación estándar. Estos dos son medidas de dispersión de una variable alrededor de la media, y la relación entre los dos es muy sencilla: la varianza (*V*) es el cuadrado de la desviación estándar (*DE*).

$$V = DE^2$$

No se trata en este espacio de explicar las formas de cálculo de la varianza, o de la desviación estándar. Para eso utilizamos los paquetes estadísticos. Se intenta explicar el sentido de estos indicadores a partir del análisis de las fórmulas que los definen. La siguiente es la fórmula que define la *varianza* o, lo que es lo mismo, la desviación estándar al cuadrado:

$$V = DE^2 = \frac{\sum_1^N (x-M)^2}{N}$$

Donde *M* es la media de los datos (*N* datos).

El concepto se puede entender como el promedio de los desvíos al cuadrado. La expresión (*X* – *M*) representa el *desvío* que presenta el dato (*X*), frente a la media general (*M*); o sea, la distancia presente entre la media y cada dato. Estos desvíos se elevan al cuadrado y se promedian (esto es, se dividen por el número total de desvíos, que es *N*). Este promedio es la varianza. Es un número siempre positivo, que se denota mediante la letra *V*.

La *desviación típica*, o *desviación estándar,* es la raíz cuadrada de la varianza. De alguna forma, esta raíz cuadrada compensa la elevación de los desvíos al cuadrado para promediarlos en el cálculo de la varianza; es en este sentido que la desviación estándar puede entenderse como una medida cercana al promedio de las distancias que tienen los datos respecto de su media. Siempre es un resultado positivo y se denota con las letras *DE* para las muestras (*SD* en inglés) y la letra griega σ (sigma) para las poblaciones.

La desviación típica se determina, siempre, en variables numéricas con niveles de medida de intervalo o de razón y se expresa en las mismas unidades de la variable y de la media. Se trata de una

medida que indica qué tan dispersos se encuentran los datos. Cuanto más pequeña sea esta, indica menor dispersión de los datos.

Error estándar de la media

El *error estándar de la media* (EEM, o SEM en inglés) es una medida que cuantifica las variaciones de la media muestral (M) alrededor de la media poblacional (μ). El EEM se estima a través de la siguiente fórmula:

$$EEM = \frac{DE}{\sqrt{n}}$$

Una observación de la fórmula del EEM indica que este relaciona la desviación estándar con el tamaño de la muestra. De esta forma, si la desviación estándar crece, crece también el EEM; por otro lado, si la muestra es grande, el EEM disminuye.

El error estándar de la media nos indica qué tanto podemos confiar en la media muestral como estimador de la media poblacional. Si el EEM es pequeño, podemos asumir que la media muestral es muy cercana a la media poblacional. Esto ocurrirá con una desviación estándar pequeña o con una muestra grande.

Coeficiente de variación

En algunas oportunidades es interesante comparar los niveles de dispersión de diferentes variables presentes en las personas de un mismo grupo, medidas en unidades diferentes. Así, si se quisiera saber qué varía más en los alumnos de una clase, si sus edades (mediadas en años) o su rendimiento en una prueba de Matemáticas, entonces, la respuesta debe buscarse con el uso del llamado *coeficiente de variación* (*CV*), que se expresa con la siguiente ecuación:

$$CV = \frac{DE}{\overline{M}} * 100$$

En la cual *DE* es la desviación estándar del grupo, *M* es la media. El coeficiente de variación se expresa en términos de porcentaje (%).

Ejemplo. En la tabla 10 se presentan los estadísticos descriptivos (media y desviación estándar) de un conjunto de cuatro pruebas (competencias en Matemáticas, Lenguaje, Ciencias Naturales y la prueba EFT) aplicadas a una muestra de estudiantes a la que nos hemos referido antes en este capítulo. En la última columna aparece el dato del coeficiente de variación para cada prueba.

Es interesante observar cómo, en principio, tenemos cuatro instrumentos con variaciones muy notorias en sus desviaciones estándar: la prueba de Lenguaje mostraba una desviación estándar de 47,84, mientras que la prueba EFT la mostraba de 9,62, que parece pequeña comparada con la anterior. Sin embargo, el cálculo de los coeficientes de variación muestra que, comparativamente, el puntaje con mayor variabilidad es, precisamente, la prueba EFT (*CV*=38,92 %).

Tabla 10. Estadísticos descriptivos de cuatro instrumentos con sus coeficientes de variación

Prueba	N	Media *M*	Desviación estándar *DE*	Coeficiente de variación (%)
Prueba de competencias en Matemáticas	202	102,68	39,13	38,10
Pruebas de competencias en Lenguaje	202	135,64	47,84	35,27
Prueba de competencias en Ciencias Naturales	202	105,26	36,33	34,52
Puntaje en la prueba EFT	231	24,71	9,62	38,92

Ejemplo: medidas de tendencia central y variación

Para el caso de las cuatro variables ficticias correspondientes a la edad de un grupo de 20 integrantes, la figura 24 muestra y compara los diferentes indicadores de tendencia central y dispersión que hemos considerado.

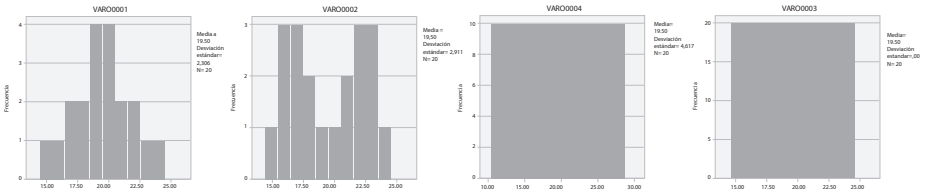
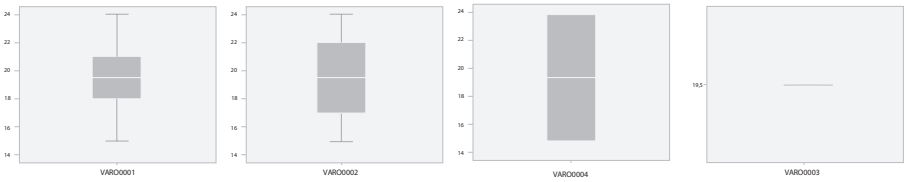
	Edad 1. Central	Edad 2. Bimodal	Edad 3. Extremos	Edad 4. Iguales
Histograma				
Media	19,50	19,50	19,50	19,50
Mediana	19,50	19,50	19,50	19,50
Moda(s)	Dos: 19 y 20	Cuatro: 16, 17, 22 y 23	Dos: 16 y 24	Una: 19,50
Diagrama de cajas y bigotes				
Rango	9	9	9	0
Rango intercuartil	3	5	9	0
Varianza	5,32	8,47	21,32	0
Desviación estándar	2,31	2,91	4,62	0
Error estándar de la media	0,52	0,65	1,03	0
Coeficiente de variación	11,84	14,92	23,69	0

Figura 24. Tendencia central y dispersión de cuatro variables de edad en una muestra de 20 personas

Puntuaciones Z

Una *puntuación Z*, o *puntuación estándar*, es la transformación de una observación que indica qué cantidad de desviaciones estándar se encuentra esta observación por encima o por debajo de la media. Así, la desviación estándar de una variable se convierte en la unidad de medida, en el patrón, de la nueva variable.

En esencia, cuando transformamos una variable en sus puntuaciones Z estamos creando una nueva variable que es idéntica, en casi todo, a la inicial. La única diferencia entre la nueva variable y la anterior son las unidades en que esta expresa. Las unidades de la nueva variable son unidades de desviación estándar de la variable original.

Volvamos a nuestro ejemplo, en el que se aplicó una prueba EFT, cuyo puntaje mínimo posible es de 0 puntos y máximo posible es de 50 puntos, a 231 estudiantes de la educación básica secundaria. La media de la prueba fue 24,71 puntos, y la desviación estándar fue de 9,62 puntos. A los puntajes originales los llamaremos ahora *puntuaciones brutas*.

Sobre esta información, ya podemos "estandarizar" el puntaje EFT. Para hacerlo, la fórmula es:

$$z = \frac{X - M}{DE}$$

Donde X es la puntuación bruta
M es la media, y
DE es la desviación estándar

La nueva variable estandarizada tiene varias propiedades muy interesantes. La primera es que la media de la nueva variable es 0. La segunda es que la nueva variable tiene desviación estándar 1. La tercera es que las unidades de la nueva variable están dadas en términos de desviación estándar de la variable original.

Así, es muy fácil identificar la posición de cada puntuación Z. Si una puntuación Z es +1, esto significa que la puntuación bruta de esa observación está a una desviación estándar por encima de la media; si esta es -1, significa que está a una desviación estándar por debajo de la media.

En el caso de nuestro ejemplo, una puntuación Z de -2, significa que el sujeto obtuvo una puntuación bruta que se ubica a dos desviaciones estándar por debajo de la media. Esto es, en términos de puntuaciones brutas:

$$X = 24.71\text{-}2(9.62) = 5{,}47 \text{ puntos}$$

Así, con un solo número, hemos podido sintetizar muchos datos acerca de la posición relativa de cada puntuación dentro de la distribución de la variable original.

Las puntuaciones Z tienen muchísimas aplicaciones en educación, y muy especialmente en el reporte de resultados de pruebas. Una de las posibilidades más interesantes es que permiten establecer comparaciones entre variables completamente diferentes. Observen por ejemplo los puntajes, expresados en puntuaciones Z, de una estudiante, Juana (datos reales y nombre ficticio), en las cuatro pruebas diferentes que hemos analizado.

Tabla 11. Puntajes estandarizados de Juana en cuatro pruebas diferentes

Prueba	*M*	*DE*	*Puntuación Z*
Prueba de competencias en Matemáticas	102,68	39,13	0,31
Prueba de competencias en Lenguaje	135,64	47,84	-0,47
Prueba de competencias en Ciencias Naturales	105,26	36,33	-0,06
Puntaje en la prueba EFT	24,71	9,62	-1,22

Una observación muy rápida de los puntajes nos permite ubicar el desempeño de Juana en las cuatro pruebas. Aparentemente a Juana le fue bien en la prueba de Matemáticas: está 0,31 desviaciones estándar por encima de la media; su desempeño en Lenguaje fue más bien pobre: 0,47 desviaciones estándar por debajo de la media. En la prueba de Ciencias, Juana se encuentra prácticamente sobre la media del total de estudiantes (apenas 0,06 desviaciones estándar por debajo). El puntaje en la prueba EFT sí que parece estar bastante por debajo de sus compañeros: ¡a 1,22 desviaciones estándar por debajo de la media!

Cómo obtener medidas de tendencia central, dispersión y puntuaciones Z en los programas

Para obtener las diferentes medidas de tendencia central y dispersión de las que se ha hablado en este capítulo, en el programa JASP, debe partirse del menú "*Descriptives*" en el directorio raíz (recuadro 7). Existen muchas opciones en este menú, especialmente de tipo gráfico. Las que hemos descrito en este capítulo aparecen en la sección "*Statistics*". Para calcular estos valores en el IBM-SPSS hay varios caminos. En el primero, debe procederse a través del menú /Analizar/Estadísticos descriptivos/Descriptivos… (recuadro 8).

Recuadro 7. Cómo obtener medidas de tendencia central y dispersión en JASP

/Descriptives
En este punto, las variables por describir deben ser pasadas a la lista "Variables". Es posible elegir una variable nominal, u ordinal, para pasarla a la lista "*Split*". En el caso en que esto se haga, todos los descriptivos se presentarán por separado, para cada valor de la variable

Statistics
- Central Tendency
  - √ Mean
  - √ Median
  - √ Mode
- Dispersion
  - √ S.E. Mean
  - √ Std. deviation
  - √ IQR
  - √ Variance
  - √ Range
- Distribution
  - √ Skewness
  - √ Kurtosis

Recuadro 8. Cómo obtener medidas de tendencia central y dispersión en IBM-SPSS

/Analizar/Estadísticos descriptivos/Descriptivos…
En este punto, las variables por describir deben ser seleccionadas y pasadas a la lista "Variables"
√ Guardar valores estandarizados como variables
Opciones
√ Media
Dispersión
√ Desviación estándar
√ Varianza
√ Rango √ Media de error estándar
Distribución
√ Curtosis √ Asimetría
Pulsar "Continuar"
Pulsar "Aceptar"

El cálculo de una puntuación Z a partir de la variable original es muy sencillo en el IBM-SPSS. En el menú /Analizar/Estadísticos descriptivos/Descriptivos… existe una casilla "Guardar valores estandarizados como variables". Si se selecciona esa casilla, el programa construirá nuevas variables para todas las seleccionadas. Estas nuevas variables contienen las puntuaciones Z de las variables originales. Muy convenientemente, los nombres de las nuevas variables son "Z —nombre de la variable original—". Para obtener diferentes medidas de tendencia central y dispersión de una variable en el IBM-SPSS puede procederse a través del menú "Explorar" (recuadro 9).

Recuadro 9. Cómo obtener más medidas de tendencia central y dispersión en IBM-SPSS

/Analizar/Estadísticos descriptivos/Explorar…
En este punto las variables por describir deben ser pasadas a la "Lista de dependientes". Es posible pasar una o varias variables nominales a otra lista, "Lista de factores"
Estadísticos
√ Descriptivos
Pulsar "Continuar"
Pulsar "Aceptar"·

Este menú ofrece y aporta multitud de descriptivos univariados bastante útiles, medidas de tendencia central y dispersión, asimetría y curtosis, etc., muchos de los cuales no han sido presentados en este capítulo, además de presentar gráficas de cajas y bigotes y de tallo y hojas. Formas en que se reportan las medidas de tendencia central y variabilidad en publicaciones científicas.

En artículos, informes y publicaciones científicas es muy frecuente que se expresen los resultados de medias y desviaciones estándar. Por el contrario, los datos de medianas y modas, rangos y rangos intercuartiles son poco frecuentes. Los datos de varianzas y puntuaciones Z, aunque se utilizan a menudo, rara vez se reportan.

Habitualmente, las medias y las desviaciones estándar se expresan en la misma línea y de seguido. Observe el siguiente párrafo:

> *The age of the high school students' oscillated between 15 and 19 years (M=16.30 years, SD=0.93); the undergraduate students between 16 and 30 years (M=20.18 years, SD=3.33), and, finally, the postgraduate students between 22 and 40 years old (M=33.83 years, SD=5.25). (Hederich, Camargo y López, 2018, p. 127)*

Una convención bastante frecuente es expresar las medias y, de corrido, las desviaciones estándar correspondientes entre paréntesis: "Las medias (con las desviaciones estándar entre paréntesis) para los Ensayos del 1 al 4 fueron 2.43 (0.50), 2,59 (1.21), 2.68 (0.39) y 2.86 (0.12), respectivamente" (APA, 2010, p. 118).

Es muy frecuente expresar medias y desviaciones estándar en tablas, en columnas sucesivas:

> *The data show that, in general, the highest means correspond to the motivation towards studying, while the lowest means correspond to the cognitive strategies. The scale with the highest mean corresponds to task value (M=5.61; SD=0.88) and the scale with the lowest value corresponds to the use of strategies of time management and study environment (M=4.58; SD=0.85).*

Table 3. Descriptive Statistics of the Motivation Strategies for Learning Questionnaire (MSLQ)

	Category	Mean	SD	Cronbach's alpha
Motivation	Intrinsic goals	5.59	0.85	.65
	Extrinsic goals	5.44	1.10	.69
	Task value	5.61	0.88	.81
	Control beliefs	5.34	0.89	.53
	Self-efficacy	5.50	0.77	.81
	anxiety	4.39	1.20	.72
Strategies	Review strategies	4.83	1.00	.62
	Elaboration strategies	4.99	0.93	.73
	Organization strategies	4.74	1.16	.69
	Critical thinking strategies	4.85	0.93	.70
	Metacognition	4.77	0.71	.69
	Study time and environment	4.58	0.85	.59
	Effort for regulation	4.84	0.94	.46
	Learning in pairs	4.71	1.10	.57
	Requesting help	4.96	0.92	.47

(Hederich et al., 2018, p. 130).

Capítulo 4

Describir relaciones entre dos variables: correlación y medidas de asociación

Presentación

Hasta este momento hemos dedicado nuestros esfuerzos a examinar todos los dispositivos conceptuales para la descripción del comportamiento de una variable; hemos presentado estadísticas *univariadas.* Iniciamos en este punto la exploración de la estadística que utilizamos para describir la forma de la relación entre dos variables, con lo que abrimos el campo de las estadísticas *bivariadas.*

En muchos proyectos de investigación social y educativa, nos interesa describir y comprender relaciones entre variables y determinar, de alguna manera, la fuerza y el sentido de estas relaciones.

Las relaciones entre variables pueden ser múltiples y de muy diferentes tipos, dependiendo de la naturaleza y el número de las variables involucradas. En relación con los tipos de relación, se pueden estudiar relaciones lineales o no lineales (cuadráticas, inversas, logarítmicas, exponenciales, polinómicas de cualquier grado, etc.). Una relación lineal entre dos variables es aquella que puede ser bien descrita por una recta.

En cuanto al número de variables involucradas, cuando buscamos relaciones entre dos variables, hablaremos de relaciones bivariadas**.** Aunque serán las que estudiaremos en este capítulo, las correlaciones bivariadas no son el único tipo de correlación posible: podemos también establecer el grado de relación entre dos variables controlando el efecto de una tercera; en ese caso hablaríamos de correlaciones parciales. Incluso, existe un procedimiento multivariante para calcular las correlaciones entre dos conjuntos de variables: en ese caso, hablaríamos de correlaciones canónicas. No abordaremos ninguno de estos temas avanzados en este capítulo, cuya naturaleza es más bien introductoria. Específicamente, examinaremos diferentes tipos de relaciones bivariadas, dependiendo de los diferentes niveles de medida de las variables involucradas.

Relaciones lineales y no lineales

Cómo graficar relaciones entre dos variables

Son múltiples las posibilidades de relación entre dos variables. Para tratar de entender esta multiplicidad, intentaremos expresarlo de forma gráfica. Para examinar, por inspección visual, la forma de relación presente entre dos variables, o la ausencia de esta relación, existe un tipo de diagrama de uso común entre variables numéricas y continuas: el llamado diagrama de dispersión.

Un *diagrama de dispersión* es una gráfica del tipo XY en la que una variable se sitúa en el eje Y, o de las ordenadas, y la otra variable en el eje X, o de las abscisas. En la medida en que cada caso presenta un valor en cada variable, el mismo quedará representado en la gráfica mediante un punto ubicado en el lugar en que se cruzan los dos valores.

En la figura 25 se muestra el diagrama de dispersión de dos variables que hemos examinado antes: el puntaje en la prueba de competencias en Ciencias Naturales y el puntaje en la prueba de Matemáticas, entre 1242 estudiantes de grado 10.°, tal y como lo genera el IBM-SPSS.

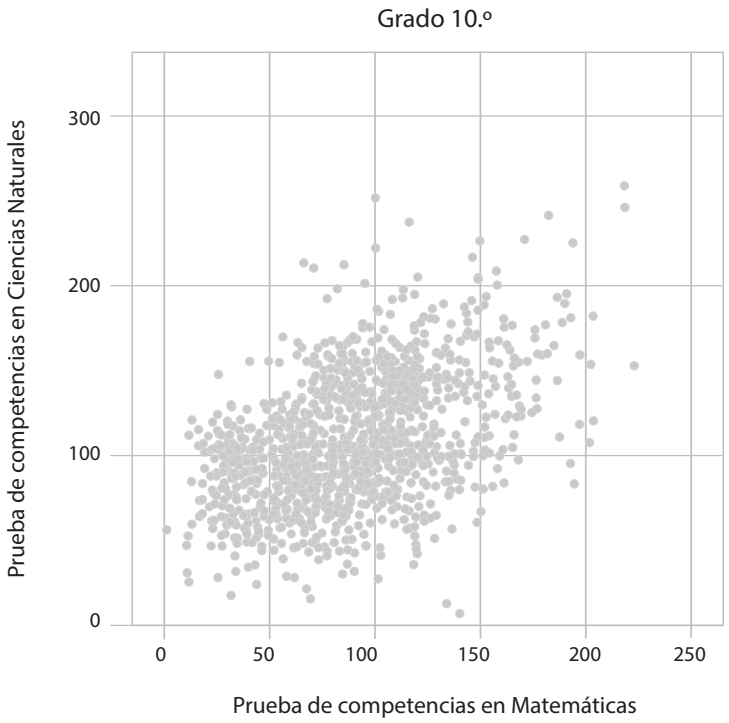


Figura 25. Diagrama de dispersión entre los puntajes de las pruebas de Ciencias y Matemáticas (n=1242)

La inspección visual de esta gráfica nos informa de la presencia de cierta relación entre estos dos puntajes. Aparentemente, hay una cierta dirección en esta relación: a mayores puntajes en una prueba se observa una cierta tendencia a mostrar mayores puntajes en la otra. Aunque la relación no es perfecta, pareciera mostrarse cierta linealidad; esto es, podríamos trazar mentalmente una línea recta que dé cuenta de esta relación, mientras que no es claro ningún otro tipo de función curvilínea que describa de mejor forma esta relación.

La presencia de esta linealidad es fundamental. No todas las relaciones que graficamos en diagramas de dispersión muestran un comportamiento lineal. Observe el ejemplo de la figura 26.

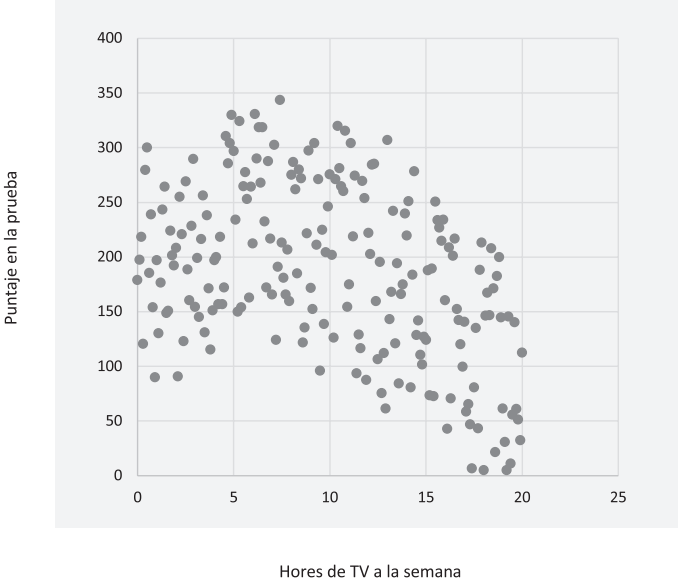


Figura 26. Diagrama de dispersión del puntaje en una prueba vs. horas diarias de TV a la semana (datos ficticios)

En la investigación social y educativa, son muchas las parejas de variables que muestran relaciones no lineales entre sí. Un ejemplo clásico está dado por la relación entre el número de horas semanales que un estudiante ve televisión en casa y sus rendimientos académicos (entendidos estos últimos en términos de notas o de puntajes en alguna evaluación estandarizada). En repetidas ocasiones los resultados que examinaron las relaciones entre estas dos variables constataron que, en principio, la relación parecía ser creciente: a mayor número de horas en la TV, mayor rendimiento; alcanzado cierto número de horas, la relación se estabilizaba y posteriormente se invertía; a partir de allí, a mayor número de horas, menor rendimiento.

La explicación de este comportamiento podría ser bastante compleja y pasa por múltiples consideraciones. Para la explicación de la primera parte, en donde se evidencia una relación creciente entre las variables, pueden mencionarse, entre otras cosas, que la TV puede ser un vehículo de acceso a la cultura y al conocimiento, o que cierto nivel de esparcimiento podría favorecer un buen desempeño académico. Incluso podríamos anotar que, en el momento histórico en el que se hacían estos estudios, se presentaba una relación entre el nivel socioeconómico y el número de horas en que se veía TV, por lo cual el nivel socioeconómico del estudiante explicaría esta relación. La explicación de la segunda parte, que indica relaciones negativas a partir de cierto número de horas es igualmente compleja y podría incluir desde un estado de desatención parental hasta una tendencia evasiva en el estudiante frente a sus labores escolares.

La *cronopsicología*, que estudia las variaciones de comportamiento a lo largo del día, nos ofrece otro ejemplo de relaciones no lineales, y es el presente entre el *nivel de activación* y el *nivel de eficiencia* del individuo en tareas cognitivas de variada complejidad. Los resultados de la investigación cronopsicológica han indicado, repetidamente, que la mayor eficiencia en tareas cognitivas

complejas se da con niveles medios de activación: no demasiado bajos, en donde se estaría casi dormido, ni demasiado altos, en donde habría una sobreexcitación. Así, la relación entre los niveles de eficiencia y los niveles de activación durante las horas de la mañana muestra un patrón no lineal: inicialmente la relación es creciente, hasta las diez horas, y a partir de ese momento empieza a ser decreciente (Testu, 1998).

En general, hablamos de *correlación* para indicar la fuerza y dirección de la relación entre dos variables. Si no se dice lo contrario, cuando se habla de correlación en realidad se está hablando de correlación lineal. Una *correlación lineal* es aquella en que la relación entre las dos magnitudes puede ser descrita por una recta; en otras palabras, es una relación en la que las dos magnitudes (X y Y) cumplen, de forma aproximada, que

$$\mathbf{Y \approx m \cdot X + b,}\ \text{siendo m cualquier número diferente de 0}$$

En la medida en que en este capítulo estaremos buscando relaciones lineales entre dos variables, el primer paso que deberemos seguir es la constatación de la linealidad de la relación entre las variables. Para realizar esta constatación deberemos siempre, como primer paso, examinar el diagrama de dispersión de las dos variables. Si esta inspección indica la posibilidad de una relación lineal, podremos proceder al cálculo de la correlación. Si, por el contrario, la inspección del diagrama de dispersión sugiere la presencia de correlaciones no lineales, no deberíamos proceder en ese sentido. Habría que buscar otro tipo de procedimientos, como por ejemplo la llamada *estimación curvilínea* en IBM-SPSS, para estimar otras formas de relación entre las variables.

La figura 27 ilustra algunas parejas de variables estrechamente correlacionadas, pero con relaciones no lineales. Si intentáramos expresar la relación entre estas variables como una relación lineal, estaríamos ignorando la naturaleza de su relación y, con seguridad, subestimando la magnitud de esta.

Figura 27. Diagramas de dispersión que indican relaciones no lineales

El diagrama de dispersión nos alerta también sobre la presencia de datos atípicos o extremos. Más adelante examinaremos el efecto de la presencia de este tipo de datos sobre las correlaciones.

Cómo obtener diagramas de dispersión en los programas

Para hacer diagramas de dispersión entre dos variables, podemos recurrir a los diferentes programas estadísticos que utilizamos, o al MS-Excel. Para dibujar diagramas de dispersión en el programa JASP puede seguirse el camino señalado en el recuadro 10. Para hacerlo en IBM-SPSS, puede recurrirse al menú /Gráficos y allí optar por "Generador de gráficos" o bien por "Cuadros de diálogo antiguos" (recuadro 11).

Recuadro 10. Cómo obtener un diagrama de dispersión en JASP

/Regression
Classical-Correlation
En este punto deben pasarse las variables que se desea examinar (pueden ser varias)
X Scatter plots

Recuadro 11. Cómo obtener un diagrama de dispersión en SPSS

/Gráficos/Cuadros de diálogo antiguos/ Dispersión/Puntos…
En este punto deberán seleccionarse los diagramas de Dispersión simple (para dos variables) o Dispersión matricial (para más de dos variables)
Pulsar "Definir"
Arrastrar las variables deseadas
Pulsar "Aceptar"

Una vez se ha constatado, mediante inspección visual del diagrama de dispersión, que podría haber una relación lineal entre las dos variables que consideramos, podemos pasar al cálculo de la medida de esa asociación.

Coeficientes de correlación y medidas de asociación

En general, el número que expresa la fuerza y el sentido de la relación entre dos variables numéricas, u ordinales, se conoce como *coeficiente de correlación*. Existen diferentes formas de determinar ese número en estadística, y la selección de la forma adecuada depende del nivel de medida de las variables involucradas.

La tabla 12 presenta los coeficientes de correlación más populares entre variables numéricas y ordinales. Estudiaremos algunos de ellos.

Tabla 12. Algunos coeficientes de correlación bivariada, para variables numéricas u ordinales

Nivel de medida	Nombre de la correlación	Símbolo*
Dos variables numéricas, de intervalo o de razón	Correlación producto-momento de Pearson	r
Dos variables ordinales	Coeficiente de correlación de rangos de Spearman	r_s, r
	Coeficiente de correlación rango-orden de Kendall	τ

*, según el manual de publicaciones APA (2016) (versión APA 6).

Independientemente del tipo de correlación que se use, hay unas características comunes a todos los coeficientes de correlación lineal:

- Se requiere siempre de dos variables, cuyos valores han sido determinados en el mismo conjunto de individuos.
- Los valores de los coeficientes de correlación varían entre -1 y 1. Ambos extremos representan correlaciones perfectas entre las dos variables y el 0 representa la ausencia de correlación.
- Una *correlación positiva* se obtiene cuando, a valores altos de una variable se presentan valores altos de la otra o, lo que es lo mismo, a valores bajos en una se presentan valores bajos en la otra. Decimos en este caso que la relación es *directa.*
- Una *correlación negativa*, o una *relación inversa*, se obtiene cuando, a valores altos en una variable se presentan valores bajos en la otra o, lo que es lo mismo, a valores bajos en una se dan valores altos en la otra.

Para el caso de variables nominales, dicotómicas o politómicas, no utilizamos la expresión "correlación", sino que nos referimos a *medidas de asociación.* Algunas de las medidas de asociación entre variables nominales más conocidas se listan en la tabla 13.

Tabla 13. Medidas de asociación entre variables nominales

Nivel de medida de las variables nominales	Nombre de la medida de asociación	Símbolo
Para variables politómicas	El coeficiente de contingencia C	*C*
	El coeficiente *V* de Cramer	*V*
Idealmente, en dos variables dicotómicas	El coeficiente Phi	Φ

Al respecto de estas medidas de asociación entre variables nominales, debe anotarse:

- Por la naturaleza de las variables nominales, en este caso no puede hablarse de relación lineal ni de correlaciones positivas o negativas.
- En general, las medidas de asociación entre variables nominales pueden variar entre 0 y 1, en donde 0 representa ausencia de relación y 1 representa relación perfecta.
- Excepción a lo anterior es el coeficiente de contingencia (*C*), que puede ser mayor que 1 en tablas grandes.

En lo que sigue se examinarán, con algún nivel de detalle, tres casos: 1) el de dos variables cuantitativas; 2) el de dos variables ordinales y 3) el de las asociaciones entre variables nominales.

Dos variables cuantitativas: el coeficiente de correlación de Pearson

El concepto

El *coeficiente de correlación de Pearson* (*r*), también conocido como coeficiente de correlación producto-momento de Pearson, es la medida numérica más conocida para determinar la relación existente entre dos variables numéricas de intervalo o razón. Puede ser entendido como una

medida del grado en el cual las parejas de datos ocupan posiciones iguales (u opuestas) dentro de sus propias distribuciones.

Como lo dijimos antes, la *r* de Pearson puede tomar cualquier valor entre -1 y 1. Una *r*=1 o *r*=-1 indican una *correlación perfecta*: las dos variables son prácticamente idénticas en el caso de *r*=1, o solo difieren en su sentido, en el caso de *r*=-1. Dicho de otra forma, en una correlación perfecta, todos los puntos se encuentran exactamente sobre la recta. En una *correlación imperfecta*, aunque existe una relación, no todos los puntos se encuentran sobre la recta. La enorme mayoría, por no decir que la totalidad, de las correlaciones que encontramos en la investigación educativa o social no son perfectas y, de hecho, difieren mucho de este ideal. En la figura 28 se ilustran diferentes grados de correlación entre dos variables cuantitativas.

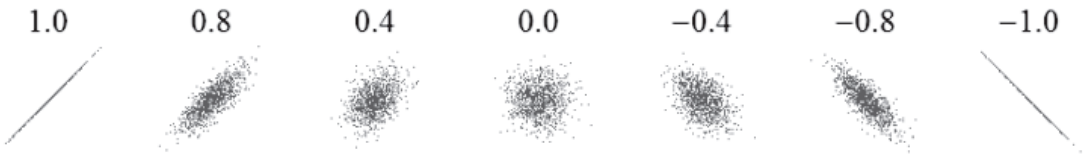


Figura 28. Correlaciones de Pearson de diferentes diagramas de dispersión

Si no existe en absoluto relación entre dos conjuntos de variables, la *r* de Pearson será cero (0) o un valor cercano. Esto indicará que la relación lineal entre las dos variables es nula, o demasiado débil para ser considerada.

La tabla 14 podría ser utilizada para interpretar los diferentes valores del coeficiente de correlación.

Tabla 14. Interpretación de los valores del coeficiente de correlación

Valor de $\|r\|$	Interpretación
$\|r\| = 1$	correlación perfecta
$0{,}8 < \|r\| < 1$	correlación muy alta
$0{,}6 < \|r\| < 0{,}8$	correlación alta
$0{,}4 < \|r\| < 0{,}6$	correlación moderada
$0{,}2 < \|r\| < 0{,}4$	correlación baja
$0{,}0 < \|r\| < 0{,}2$	correlación muy baja
$r = 0$	correlación nula

Como en otros temas, no explicaremos las fórmulas ni los pasos necesarios para calcular el coeficiente de correlación de Pearson. Una aproximación intuitiva para calcularlo podría hacerse a partir del cálculo del *promedio de los productos cruzados de las puntuaciones Z de las dos variables involucradas*. Explicaremos brevemente el sentido de esta definición.

Como se recordará, las puntuaciones Z de las dos variables involucradas están dadas en unidades de desviación estándar. Tenemos tres posibles casos:

- Si un caso tiene puntajes altos en las dos variables, sus puntuaciones Z serán positivas y, por tanto, sus productos serán positivos. Si, por el contrario, tiene valores bajos en las dos variables, las puntuaciones Z serán negativas y sus productos positivos. Así, las sumas de los productos serán positivos y su promedio hará que el valor sea positivo y mayor que 1. En ese caso, tendremos un coeficiente de correlación positivo y alto.
- Por otro lado, si los valores altos en una variable coinciden, consistentemente, con valores bajos en la otra, los productos serán negativos y su suma también será negativa, por lo que el coeficiente de correlación será negativo.
- Por último, si a valores altos en una variable se le aparejan, erráticamente, valores altos, medios o bajos en la otra, los productos tendrán signos erráticos y su suma no se alejará mucho del 0. En este caso, el coeficiente será cercano a 0.

La predicción de una variable por otra

Es intuitivamente claro que cuando dos variables numéricas están correlacionadas es posible, en alguna medida, "predecir" los valores de una de ellas a partir de los valores de la otra. Si la correlación fuera perfecta, esto sería muy sencillo: existe una recta que pasa por todos los puntos y se trataría de determinar esa recta y reemplazar el valor en una variable para obtener el valor exacto en la otra. Obtendríamos el valor de la segunda variable con absoluta precisión.

¿Y si la relación entre las dos variables es imperfecta? La verdad es que podríamos también hacerlo, aunque ahora nuestra predicción sería, igualmente, imperfecta, como imperfecta es la correlación.

Para hacer esta predicción, podemos interpretar el coeficiente *r* de Pearson en términos del grado de variabilidad de una variable (Y) que puede ser explicado por la otra (X). Más específicamente, se puede demostrar —aunque no lo haremos en este espacio—, que el coeficiente *r* de Pearson es igual a la raíz cuadrada de la proporción de la variabilidad de Y que es explicada por X. En otras palabras,

r^2 es la proporción de la variabilidad total de Y que es explicada por X

Este valor (r^2) es muy importante, ya que nos da una medida precisa de la capacidad de predicción de una variable por otra. Es conocido como *coeficiente de determinación*.

Así, el valor de r^2 nos permite valorar la verdadera magnitud del coeficiente *r* de Pearson. Obsérvese la tabla 15, en donde aparecen algunos cuadrados de valores entre 0 y 1. Por supuesto, si r=0, la capacidad explicativa de una variable por la otra es 0 %. En el otro extremo, si *r*=1, tenemos una relación perfecta con una capacidad explicativa del 100 %. Ahora, para juzgar los valores intermedios, debe observarse que la diferencia entre una *r*=,1 y una *r*=,2, además de que la segunda es el doble de la primera, es que la proporción de variabilidad explicada por *r*=,2 es *cuatro veces* la explicada por *r*=,1.

Tabla 15. Coeficientes de correlación (r) y coeficientes de determinación (r^2)

r	r^2	**Proporción de variabilidad explicada (%)**
0	0	0
,1	,01	1
,2	,04	4
,3	,09	9
,4	,16	16
,5	,25	25
,6	,36	36
,7	,49	49
,8	,64	64
,9	,81	81
1	1	100

Esto nos cambia de forma muy importante la interpretación del valor del coeficiente *r*. Obsérvese que un *r*=,5 representa una explicación de, apenas, el 25 % de la variabilidad. Para que logremos una explicación del 50 %, necesitamos un coeficiente de correlación de Pearson levemente mayor que 7. Volveremos sobre el coeficiente de determinación en el próximo capítulo, dedicado a regresión lineal.

La significación estadística de r

Como se trata habitualmente, el coeficiente de correlación es un *estadístico descriptivo*, dado que muestra el grado y la dirección de la relación lineal entre dos variables numéricas, medidas en una muestra de individuos. Así se aborda en este capítulo.

Ahora, en muchos proyectos de investigación interesa saber si la relación encontrada en la muestra es generalizable a la totalidad de la población o si, por el contrario, es más bien una característica de la muestra específica. En estos casos, la correlación no se entiende como estadístico descriptivo, sino como una prueba de hipótesis. Este es un tema diferente que trataremos a partir del capítulo 7, donde iniciamos la estadística inferencial.

Por ahora, baste decir que, en ciertas condiciones, es posible calcular la probabilidad de haber obtenido un determinado coeficiente de correlación por efecto del puro azar. Esto es lo que llamaremos significación estadística (sig.) o *valor p*; es decir, la probabilidad de haber obtenido una correlación dada en una muestra, al estar ausente esa correlación en la población. Cuanto menor sea el nivel de significación, más probable es que la relación encontrada no sea fruto del azar y pueda ser verificada en la población completa.

Así, cuando calculamos y reportamos la significación estadística estamos también examinando qué tan fuerte y conclusiva es la relación que estamos estudiando y cuál es el grado de generalidad de la relación encontrada en la muestra a la totalidad de la población. Por esta razón, el nivel de significación estadística dependerá del tamaño de coeficiente de correlación y del tamaño de la muestra: mayores coeficientes de correlación en muestras más grandes habitualmente serán más significativos.

Existen ciertos niveles de significación estadística que habitualmente se aceptan en la investigación educativa y social. Tradicionalmente, el nivel de significación más bajo aceptado es $p<,05$, que significa una probabilidad menor al 5 % de haber obtenido ese resultado por azar. También se identifican, como valores de referencia, $p<,01$ (menor que 1 %) o, mejor aún, $p<,001$ (menor que 0,1 %). Estos valores se identifican como un reducto de otras épocas, en donde debía procederse al cálculo de la significación a través de tablas. Actualmente, los programas estadísticos calculan y reportan los valores exactos de la significación.

Como veremos más adelante, para calcular estos niveles de significación, debemos asegurar, previamente, ciertas condiciones, o "supuestos", sobre la naturaleza de las variables, sus distribuciones y sus relaciones. En este contexto estamos tratando la correlación como estadístico descriptivo, por lo que no estamos examinando estos supuestos. Esto nos señala la importancia de interpretar con extrema reserva estos niveles de significación.

En la medida en que los programas estadísticos reportan automáticamente el nivel de significación para cada correlación, hemos preferido hacer esta breve, e incompleta, introducción al tema de la significación. El tema se desarrollará con mucho mayor detalle a partir del capítulo 7.

Sobre la interpretación de las correlaciones

Correlación es covariación, no causalidad

Cuando existe una correlación importante entre dos variables es tentador concluir que hay una relación de causalidad entre ellas; esto es, que una es la causa de la otra. Ello, aunque sabemos que es una tendencia natural en los humanos es, en general, un grave error en la investigación educativa y social.

En efecto, no puede olvidarse que cuando hablamos de correlación hablamos, en realidad, de covariación; es decir, de una variación consistente, relativamente coordinada entre dos variables, no más —ni menos—. Así, la presencia de una alta correlación entre dos variables, digamos X y Y puede ser interpretada en, al menos, cuatro formas: 1) X es la causa de Y; 2) Y es la causa de X; 3) existe una tercera variable, Z, que explica a X y a Y; y 4) la correlación entre X y Y se dio únicamente por azar (en este caso decimos que es una correlación *espuria*).

Considérese, por ejemplo, el caso del hallazgo de una correlación negativa entre dos variables: el puntaje en pruebas estandarizadas (X) y la cantidad de tiempo que dedican los padres a hacer las tareas con sus hijos (Y). Los resultados han mostrado una correlación negativa y significativa entre estas variables: a mayor tiempo dedicado por los padres a hacer la tarea con los hijos, menor es el rendimiento de los muchachos en pruebas. ¿Cómo interpretar esta correlación?

Una primera interpretación podría dictar que los padres muestran una tendencia a "hacer" las tareas de los hijos y, en esa medida, les quitan oportunidades valiosas de aprendizaje. En ese sentido, Y es la causa de X. Sin embargo, podríamos ensayar una segunda interpretación, que indicaría que los padres solo ayudan a los hijos que presentan dificultades en la materia, mientras que a los hijos que presentan desempeño suficiente o destacado no les dedican ese tiempo; esto es, ¡X es la causa de Y!

Todavía podríamos inventar una tercera explicación, aunque podría ser un poco descabellada. Los padres que tienen tiempo para ayudar a sus hijos con las tareas escolares disponen de ese tiempo por una situación crónica de desempleo. Esta situación afecta el poder adquisitivo de la familia e introduce un estrés importante en los estudiantes que afecta su desempeño académico. En este caso, la situación socioeconómica de la familia (Z) está en el origen, tanto de X como de Y.

Por supuesto que siempre es posible que una correlación, por alta que sea, se haya presentado por efecto del azar; esto es, que sea una correlación espuria. En este hipotético caso, al seleccionar otra muestra, mejor elegida y de mayor tamaño, se observaría que la correlación encontrada ya no está presente. Este punto, aunque siempre es posible, es el que es efectivamente controlado con la selección de la muestra, el cálculo de los niveles de significación (valores *p*) y la consideración de otros aspectos relacionados, tales como la potencia de la prueba estadística y las mediciones del tamaño del efecto. Hablaremos de esto, en detalle, más adelante.

En general, una vez se encuentran correlaciones, una buena práctica para la investigación indica la necesidad de ser cuidadoso elaborando diferentes interpretaciones que planteen diferentes sentidos de la causalidad y que expliquen lo observado por la posible influencia de variables ocultas. La aseveración de una relación causal solo puede ser establecida, con cierto grado de seguridad, a partir del diseño de experimentos verdaderos.

El efecto de los datos extremos

Es muy importante el efecto que puede tener la presencia de datos extremos sobre el cálculo de las correlaciones de Pearson. Observe el siguiente caso. Para el presente ejemplo, se seleccionaron los primeros 20 casos presentes en la base de datos que hemos trabajado, en dos variables: edad y puntaje en la prueba EFT. El cálculo del coeficiente *r* de Pearson sobre esta pequeña submuestra nos indica un valor negativo y no significativo entre estas dos variables *r*=-,286 *p*=,222. La figura 29a muestra el diagrama de dispersión en estos 20 casos.

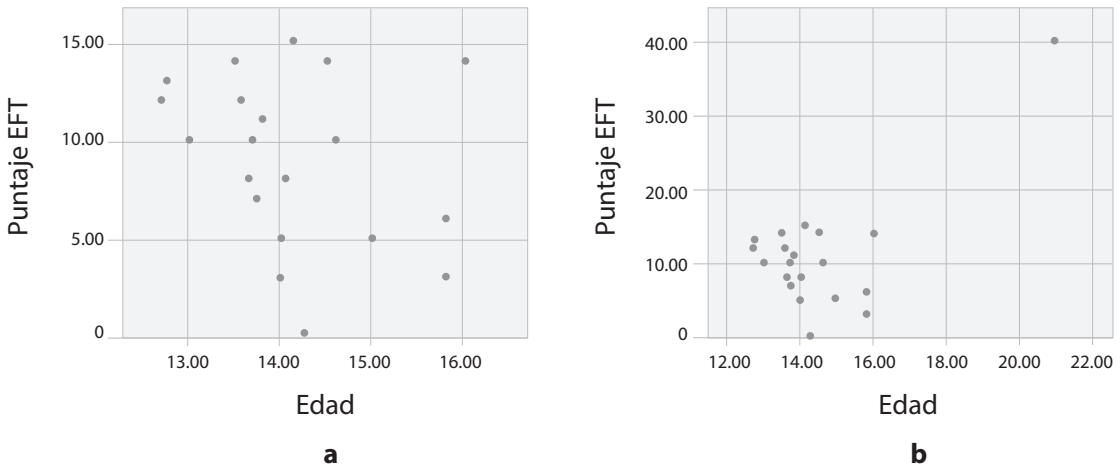


Figura 29. Dispersión de las variables Puntaje EFT y edad.

Nota: a) en 20 casos r=-,286 p=,222; b) añadiendo un caso con valores extremos r=,644 p=,002.

Como un segundo paso, a esta pequeña base se añadió un nuevo caso, que presenta valores extremos en las dos variables: un imaginario estudiante de 21 años y un EFT de 40 (estos dos valores son perfectamente posibles). La gráfica de dispersión (figura 29b) muestra con claridad este nuevo caso. Ahora, el cálculo de la correlación en esa nueva base, de 21 sujetos, muestra una correlación muy diferente, positiva, alta y muy significativa r=,644 p=,002. Esto nos muestra que el valor de la r de Pearson puede verse muy influido por la presencia de datos extremos.

Debemos tener en cuenta dos puntos para valorar este efecto de los datos extremos. Primero, debe anotarse que este efecto se presenta, de manera particular, en muestras pequeñas. Si, en vez de tener 20 casos, tuviéramos 200, el efecto del caso adicional, aunque perceptible, sería mucho más ligero. Si tuviéramos 2000 sería casi imperceptible.

Segundo, debe mencionarse que este efecto se presenta, de manera particular, en variables numéricas, en las cuales se calcula el coeficiente r de Pearson. En estas es mucho más sencillo plantear valores mucho más alejados de la nube de puntos. En variables ordinales, por otro lado, los casos "extremos" no pueden ser tan extremos, por cuanto, como máximo, identificarían un nuevo rango adicional.

En términos generales, el diagrama de dispersión permite identificar fácilmente los casos extremos. En presencia de este tipo de datos es preferible eliminar los extremos para obtener una visión más ajustada de las relaciones entre las variables.

Errores por restricción del rango

A grandes rasgos, si existe una correlación apreciable entre dos variables, una restricción del rango de cualquiera de ellas tendrá, como efecto, la disminución del coeficiente de correlación. Considere, por ejemplo, el caso de la figura 30.

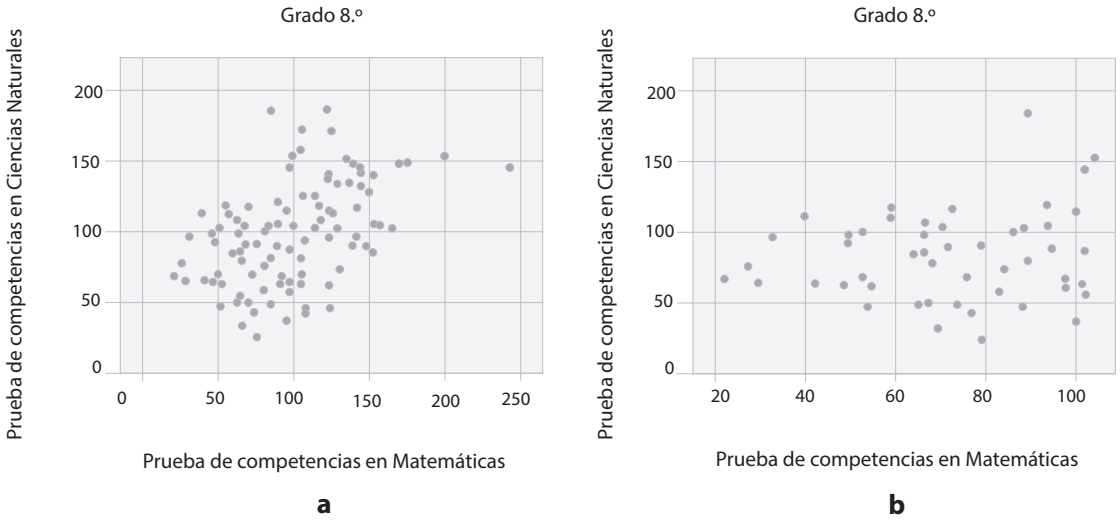


Figura 30. Efectos de la restricción del rango

Nota: a) diagrama de dispersión con rango completo; b) diagrama de dispersión con rango restringido.

El primer diagrama de dispersión (figura 30a) muestra el cruce entre el puntaje de la prueba de Matemáticas y el de Ciencias Naturales para 97 estudiantes de grado 8.º. Examinada la correlación en el rango completo, se verifica una correlación positiva, moderada y muy significativa r=,469 p<,001.

El segundo diagrama de dispersión (figura 30b) muestra el cruce de las mismas variables, pero se ha restringido el rango del puntaje de Matemáticas a aquellos puntajes menores a 100 puntos. El efecto de esta restricción, además de reducir el tamaño de la muestra (n=51), es reducir también la correlación entre las variables; la nueva correlación sigue siendo positiva, pero ahora es débil y no significativa r=,148 p=,299.

Estas restricciones en el rango pueden aparecer no solo en el momento de analizar la información, sino en el instante mismo del diseño del proyecto y de la muestra. En ese sentido, la presencia de este efecto debe alertar al investigador sobre la interpretación de correlaciones nulas o muy bajas. Es posible que las correlaciones débiles sean explicables por restricciones inadvertidas en los rangos, y no por ausencia de relaciones en sí mismas.

Cómo obtener los coeficientes de correlación en los programas

Para obtener una correlación en el programa JASP se debe proceder, en el directorio principal, a través de la opción "Regression". Las opciones recomendadas se muestran en el recuadro 12. Para obtener el coeficiente de correlación de Pearson en el IBM-SPSS, se procederá a través del menú "Correlacionar" (recuadro 13). Los dos programas generan la matriz de correlaciones de todas las variables seleccionadas en todos los tipos de correlación solicitados.

Recuadro 12. Cómo obtener una correlación en JASP

/Regression/Classical.Correlation/

En este punto se deben pasar las variables que se desean correlacionar a la lista "Variables" y seleccionar el tipo de correlación deseado:

√ Pearson´s *r*

Additional options

Report significance (estará seleccionada por defecto)

Plots

√ Scatter plots

Recuadro 13. Obtener una correlación en IBM-SPSS

/Analizar/Correlacionar/Bivariadas…

En este punto usted debe incluir todas las variables que desea correlacionar y seleccionar el tipo de correlación que desea entre tres disponibles: Por defecto, estará seleccionada la correlación de Pearson

√ Pearson

Pulsar el botón "Aceptar"

Ejemplo: la relación entre tres puntajes numéricos de tres pruebas

Se tiene información acerca de los resultados obtenidos por un amplio grupo de estudiantes de grado 10.º= en Bogotá (N=1242) en tres pruebas diferentes, aplicadas por la Secretaría de Educación de Bogotá, y que miden competencias en las asignaturas de Matemáticas, Lenguaje y Ciencias Naturales.

Primer paso: examinar los diagramas de dispersión

El primer paso es examinar los diagramas de dispersión de estas variables, dos a dos para la detección de relaciones no lineales o casos extremos. La figura 31 presenta el diagrama de dispersión matricial de las tres variables.

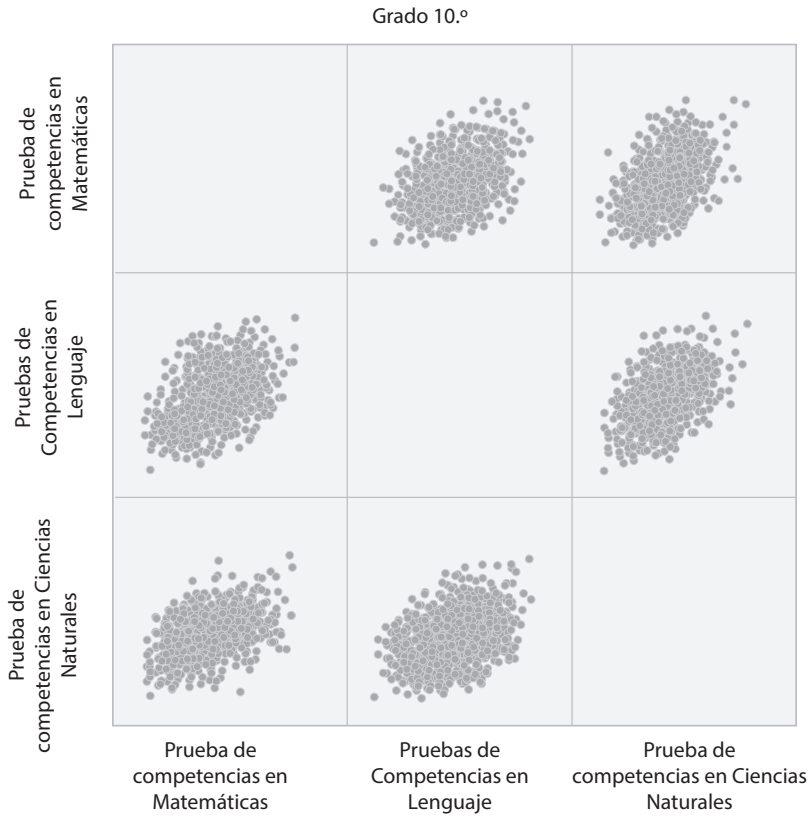


Figura 31. Diagrama de dispersión matricial de tres variables

Como se observa, la diagonal de esta matriz aparece vacía. Esto es natural, ya que la diagonal representa el cruce de cada variable consigo misma. De ser representada, aparecería una recta que indica la relación perfecta que se da entre una variable y ella misma. Para interpretar este diagrama basta con examinar los gráficos de dispersión por encima, o por debajo de la diagonal, dado que están perfectamente reflejados. La inspección visual muestra que en los tres casos parecen constatarse relaciones lineales y directas entre las variables implicadas. Este tipo de diagramas, aunque son de uso muy frecuente, casi nunca se reportan en las publicaciones científicas.

Segundo paso: calcular los coeficientes de correlación

El resultado arrojado por el SPSS sobre los coeficientes de correlación de Pearson se presenta en la tabla 16.

Tabla 16. Salida del SPSS sobre correlaciones de Pearson

		Prueba de competencias en Matemáticas	Prueba de competencias en Lenguaje	Prueba de competencias en Ciencias Naturales
Prueba de competencias en Matemáticas	Correlación de Pearson	1	,422**	,488**
	Sig. (bilateral)		,000	,000
	N	1242	1242	1242
Prueba de competencias en Lenguaje	Correlación de Pearson	,422**	1	,457**
	Sig. (bilateral)	,000		,000
	N	1242	1242	1242
Prueba de competencias en Ciencias Naturales	Correlación de Pearson	,488**	,457**	1
	Sig. (bilateral)	,000	,000	
	N	1242	1242	1242

** La correlación es significativa en el nivel 0,01 (bilateral).

Como en el diagrama matricial de dispersión, en este caso tenemos una matriz de correlaciones, cuya diagonal presenta las correlaciones perfectas de cada variable consigo misma. Así, la matriz es perfectamente simétrica, por lo que es suficiente tomar su mitad, por encima —o por debajo—, de la diagonal. En las publicaciones científicas, cuando se presentan correlaciones en forma de matriz, siempre se omite alguna de las mitades de la matriz a fin de no repetir los datos.

Tercer paso: interpretar y expresar los resultados

Como se observa, en cada cruce de variables aparecen tres filas tituladas "Correlación de Pearson", "Sig.(bilateral)" y "N". La primera contiene los coeficientes de correlación de Pearson, la segunda contiene el nivel de significación correspondiente a ese coeficiente de correlación, o su valor *p*, y la tercera contiene el número de casos sobre el cual se calculó cada correlación.

Un valor *p* de "0,000" (que, en realidad no es 0, aunque pudiera parecer), tal y como aparece en todos los cruces de nuestra matriz, significa que la probabilidad de haber obtenido ese valor como un resultado del azar es tan baja, que es de menos de 0,001 (0,1 %); esto es: ínfima. En estas condiciones diremos que la correlación es estadísticamente significativa. El valor de la significación estadística también está señalado al lado del coeficiente de correlación, mediante los pequeños asteriscos que lo acompañan. Si el valor *p* correspondiente es menor que ,01 ($p<,01$), se anotan dos asteriscos; si $,01<p<,05$, se anotará un asterisco y representa una relación significativa, aunque en menor grado; si $,05<p$ se dice que la relación no es estadísticamente significativa.

En muchas publicaciones, cuando se presentan los coeficientes de correlación, solo se muestran los asteriscos que indican los niveles de significación. En general, cuando es posible, resulta decididamente recomendable presentar todos los resultados en el texto y no repetirlos en una tabla. Una forma de hacerlo, para nuestro ejemplo, podría ser como sigue:

> *Los resultados indican una situación general de correlaciones moderadas y muy significativas. El mayor valor del coeficiente de correlación de Pearson está presente entre las pruebas de Matemáticas y Ciencias r=,488 p<,001; le siguen de cerca las correlaciones entre Ciencias Naturales y Lenguaje r=,457 p<,001 y entre Lenguaje y Matemáticas r=,422 p<,001.*

Dos variables ordinales: los coeficientes de correlación de Spearman y Kendall

El concepto

En situaciones en las que tenemos variables ordinales, no es adecuado el uso del coeficiente de correlación de Pearson. En estos casos, debemos proceder con el coeficiente de correlación r_s de Spearman, también llamado coeficiente ρ (*rho*) de Spearman, o bien con el coeficiente de correlación τ (tau) de Kendall. Estos dos métodos son llamados métodos de correlación no paramétrica, en oposición a la correlación paramétrica de Pearson. Por su naturaleza, son poco sensibles a la presencia de valores extremos.

El *coeficiente de correlación de Spearman*, también llamado ρ (*rho*) de Spearman, simbolizado r_s, puede ser utilizado en los casos en que tenemos una variable métrica y una ordinal, o dos variables ordinales. Este coeficiente parte del ordenamiento de los valores de cada variable involucrada y la asignación de rangos a cada caso. Por esta razón, este coeficiente se conoce también como coeficiente de correlación por rangos de Spearman.

La interpretación de este coeficiente es similar a la de Pearson: r_s varía entre -1 y 1. Valores negativos indican asociaciones negativas, o inversas, mientras que valores positivos indican relaciones positivas o directas. Un valor cercano a 0 indica una correlación nula, o muy baja, entre las variables.

El coeficiente de Spearman tiene un amplio uso en la investigación educativa y social en situaciones con variables ordinales o en situaciones en las que las variables numéricas no se distribuyen de forma normal.

Por su parte, *el coeficiente tau de Kendall*, también conocido como coeficiente tau-B de Kendall, simbolizado como τ, es bastante similar a los dos anteriores, pues presenta el mismo rango y la misma interpretación que el de Pearson y el de Spearman. Sin embargo, dentro de una tendencia general que muestra grandes similitudes entre los coeficientes, una característica notable del coeficiente tau de Kendall es que, en aquellas situaciones en las que se analizan las asociaciones lineales sin la presencia de valores atípicos, reporta valores un poco más bajos con respecto a los coeficientes de Spearman y Pearson, lo cual lo hace un poco más conservador.

Este no significa que el coeficiente de Kendall sea menos preciso que los otros dos. De hecho, este coeficiente es preferido por muchos investigadores, frente al de Spearman, por el hecho de que su distribución tiende más rápidamente a la distribución normal. A pesar de ello, el uso del

coeficiente τ de Kendall es bastante menos frecuente que el de Spearman en la investigación social. En general, los resultados de los dos coeficientes, el de Spearman y el de Kendall resultan ser muy cercanos. Cualquiera de los dos resultará apropiado si tenemos variables ordinales.

Cómo obtener los coeficientes de correlación en los programas

El cálculo de los coeficientes de correlación de Spearman y Kendall se hace en el mismo menú que el de Pearson. En todos los casos, el coeficiente *r* de Pearson se encuentra preseleccionado, mientras que los de Spearman y Kendall deben ser marcados. Para obtener una correlación en el programa JASP se debe proceder, en el directorio principal, a través de la opción "Regression". Las opciones recomendadas son las que se presentan en el recuadro 14. Para obtener el coeficiente de correlación de Pearson en el IBM-SPSS, se procederá a través del menú "/Analizar/Correlacionar" (recuadro 15). Los dos programas generan la matriz de correlaciones de todas las variables seleccionadas en todos los tipos de correlación seleccionados.

Recuadro 14. Cómo obtener correlaciones no paramétricas en JASP

/Regression/Classical. Correlation
- En este punto se deben pasar las variables que se desean correlacionar a la lista "Variables" y seleccionar el tipo de correlación deseado:
- √ Spearman´s rho
- √ Kendal´s tau-b
- Additional options
  - √ Report significance (estará seleccionada por defecto)
- Plots
  - √ Scatter plots

Recuadro 15. Cómo obtener correlaciones no paramétricas en IBM-SPSS

/Analizar/Correlacionar/Bivariadas…
- En este punto se deben incluir todas las variables que se busca correlacionar y seleccionar el tipo de correlación que se desea entre tres disponibles. Por defecto, estará seleccionada la correlación de Pearson
- √ Tau-b de Kendall
- √ Spearman

Pulsar el botón "Aceptar"

Ejemplo: relaciones entre evaluaciones de diferentes maestros

Contamos con las calificaciones cualitativas, de cada uno de los maestros de Matemáticas, Lenguaje y Ciencias, sobre 114 estudiantes de grado 10.º, en colegios públicos de la ciudad de Bogotá. El rendimiento individual de estudiantes en cada área fue valorado por el maestro del área en una escala ordinal de cuatro puntos, a saber, 1 "deficiente", 2 "aceptable", 3 "superior" y 4 "excelente".

Para el presente caso, un diagrama de dispersión no resulta muy útil. Los puntos quedan unos sobre los otros, lo que dificulta examinar la relación. Una tabla de contingencia puede resultar

mucho más útil para examinar las relaciones entre variables ordinales, si bien esta debería hacerse para cada par de variables por separado. En la tabla 17, se muestra el cruce entre la evaluación del maestro de Matemáticas y la evaluación del maestro de Ciencias Naturales.

Tabla 17. Tabla de contingencia entre las evaluaciones de los maestros de Matemáticas y Lenguaje

		Evaluación del Maestro de Lenguaje				Total
		1 Deficiente	2 Aceptable	3 Superior	4 Excelente	
Evaluación del maestro de Matemáticas	1 Deficiente	5	15	3	1	24
	2 Aceptable	6	30	25	1	62
	3 Superior	2	8	6	3	19
	4 Excelente	1	2	5	1	9
Total		14	55	39	6	114

a. Grado = 2 Grado 10.°

Los resultados mostrados en esta tabla parecieran sugerir la presencia de una relación lineal y directa, más bien leve. Esto podría ser visible en el hecho de que, en los extremos opuestos (excelente/deficiente o viceversa) pareciera haber muy pocos casos, mientras que en la diagonal parecieran agruparse un mayor número.

La matriz de la tabla 18 muestra las correlaciones de Spearman y Kendall entre las tres variables, con sus respectivos niveles de significación, tal y como es presentada por el JASP.

Tabla 18. Matriz de correlaciones de Spearman y Kendall entre las evaluaciones de los diferentes maestros tal y como es presentada por JASP

Variable		em_mat	em_len	em_cna
Evaluación del maestro de Matemáticas (em_mat)	Spearman's rho	—		
	p-value	—		
	Kendall's Tau B	—		
	p-value	—		
Evaluación del maestro de Lenguaje (em_len)	Spearman's rho	0,258	—	
	p-value	0,006	—	
	Kendall's Tau B	0,232	—	
	p-value	0,005	—	
Evaluación del maestro de Ciencias (em_cna)	Spearman's rho	0,140	0,032	—
	p-value	0,138	0,732	—
	Kendall's Tau B	0,123	0,030	—
	p-value	0,136	0,714	—

Como se observa, aunque en filas y columnas se listan las tres variables correlacionadas, no se presentan los resultados de la diagonal, que representarían las correlaciones de cada variable consigo misma.

Los resultados muestran un panorama general de correlaciones bajas o muy bajas. Lo primero que vale la pena observar es que los dos tipos de correlación, de Spearman y de Kendall, producen resultados muy similares, tanto en sus coeficientes como en sus niveles de significación. Lo segundo es que, consistentemente, todas las correlaciones de Kendall son levemente más bajas que las correspondientes de Spearman. Por último, debe notarse que, aunque prácticamente no hay diferencias entre los niveles de significación asociados con cada una de ellas, las significaciones de Kendall son un poco más bajas que las de Spearman.

En este espacio hemos presentado las dos correlaciones, de Spearman y Kendall, con propósitos eminentemente pedagógicos. En situaciones reales, recomendamos elegir solo una de ellas. El siguiente texto es un ejemplo de cómo se pueden expresar los resultados, utilizando las correlaciones de Kendall.

> *La mayor correlación se presenta entre la evaluación del maestro de Matemáticas y la del maestro de Lenguaje, si bien su valor puede ser considerado bajo, aunque estadísticamente significativo* τ=*,232* p=*,005. La correlación entre las evaluaciones de Lenguaje y Ciencias es prácticamente nula* τ=*,030* p=*,714, así como la correlación entre las evaluaciones de los maestros Matemáticas y Ciencias* τ=*,123* p=*,136. Las evaluaciones obtenidas por los estudiantes en las diferentes materias no parecen estar muy relacionadas entre sí.*

Medidas de asociación entre variables nominales

Aspectos conceptuales

Cuando nos enfrentamos al caso de variables estrictamente nominales, debemos recordar que para estas variables no es posible definir un orden estricto. En esta condición, es imposible hablar de relaciones directas —o positivas— e inversas —o negativas—.

En consonancia con esta característica, los coeficientes de correlación entre variables nominales, que algunos prefieren denominar como "medidas de asociación", usualmente varían entre 0 y 1, donde un valor de 0 representa la ausencia de relación y un valor de 1 representa una relación perfecta.

Las medidas de asociación entre variables nominales de uso más común son:

- El *coeficiente de contingencia C*. Medida de asociación basada en la distribución chi-cuadrado (χ^2). El valor de *C* es mayor o igual que cero y el valor 0 significa la ausencia de asociación, mientras que los valores cercanos a 1 indican su presencia. Lamentablemente, el máximo valor posible para *C* no es 1 y ello depende del número de celdas en la tabla, lo que hace muy difícil comparar los resultados de dos tablas con diferente número de filas y columnas.
- El *coeficiente Phi* (**Φ**). Medida de asociación basada en χ^2. Solo puede ser utilizado para el caso de tablas 2x2. El valor del coeficiente Phi coincide con el coeficiente *r* de Pearson y va-

ría entre -1 y 1. En tablas de dimensión mayor que 2x2, el coeficiente Phi puede ser mayor que 1 en valor absoluto.

- El *coeficiente V de Cramer*. Medida de asociación basada en χ^2 que corrige algunos de los problemas del coeficiente de contingencia *C*. El valor de *V* varía entre 0 y 1, en donde el valor 0 significa la ausencia de asociación, mientras que los valores cercanos a 1 indican su presencia. La *V* de Cramer puede alcanzar un valor de 1 en tablas de cualquier dimensión.

Aunque se han presentado estos tres coeficientes con propósitos de ilustración, se recomienda el uso del coeficiente *V* de Cramer, en gracia a su facilidad de interpretación, además de la versatilidad de su uso. En tablas de 2x2 este coincide con el coeficiente Phi. Para la interpretación del valor de la *V* de Cramer, debe mencionarse que el valor específico cambiará dependiendo del número de *grados de libertad* de la tabla de cruce. La tabla 19 presenta la interpretación de los valores de la V de Cramer, dependiendo de los grados de libertad de la tabla de cruce.

Tabla 19. Interpretación de los valores de Phi (ϕ) y V de Cramer dependiendo de los grados de libertad de la tabla

Medida de asociación	gl* (df)	Pequeño	Mediano	Grande
ϕ y *V* de Cramer	1	0,10	0,30	0,50
V de Cramer	2	0,07	0,21	0,35
	3	0,06	0,17	0,29
	4	0,05	0,15	0,25
	5	0,04	0,13	0,22

* Los grados de libertad (gl) dependen de la dimensión de la tabla. En una tabla de dimensión f*c, los grados de libertad serán gl= (f-1)*(c-1)

Fuente: a partir de Kim (2017) y de Goss-Sampson (2019).

Existen otros coeficientes de correlación para algunas parejas especiales de variables, una de las cuales es dicotómica. El cálculo de estos coeficientes no está disponible en los programas que utilizamos. Entre ellos, vale la pena mencionar los siguientes:

- El *coeficiente de correlación tetracórica* (r_t). Se utiliza cuando las variables con las que trabajamos han sido dicotomizadas de manera artificial. Es más apropiado emplear el coeficiente Phi (Φ) cuando las variables son estrictamente dicotómicas, y recurrir a r_t cuando las variables, siendo originalmente continuas, aparecen dicotomizadas.
- El *coeficiente de correlación biserial puntual* ($r_{b,}\, r^b$). También llamado coeficiente de correlación punto biserial, se utiliza cuando se busca la correlación entre dos variables: una de ellas medida en escala numérica y la otra en escala nominal dicotómica.
- El *coeficiente de correlación rango-biserial* (r_{rankb}). Derivado del coeficiente de Spearman, este coeficiente establece la relación entre dos variables, una de ellas dicotómica y la otra con un nivel de medida ordinal. Puede ser definido como la proporción de pares favorables a la hipótesis menos la proporción de pares no favorables. Se utiliza para estimar el tamaño del efecto en algunas pruebas de hipótesis.

Cómo obtener las medidas de asociación en los programas

Para el cálculo de estas pruebas, puede procederse, en JASP, de la forma presentada en el recuadro 16; para IBM-SPSS, se muestra en el recuadro 17.

Recuadro 16. Cómo obtener las medidas de asociación en JASP

/Frecuency/Contingency Tables
En este punto, debe seleccionarse la primera variable y pasarla a la lista "Rows" y la otra a la lista "Columns"
Nominal
√ Phi and Cramer's V

Recuadro 17. Cómo obtener las medidas de asociación en IBM-SPSS

/Analizar/Estadísticos descriptivos/Tablas cruzadas…
En este punto, debe seleccionarse una variable y pasarla a la lista "Filas" y la otra para pasarla a la casilla "Columnas"
En el botón "Estadísticos"
Nominal
Phi y V de Cramer
pulsar el botón "Continuar"
Pulsar el botón "Aceptar"

Ejemplos: asistencia al preescolar y cambio de colegio por género.

Para la totalidad de los estudiantes de secundaria en un colegio público de la ciudad de Bogotá (n=231), se formularon preguntas acerca de su género, su asistencia al preescolar y su matrícula en colegios diferentes al actual.

La tabla 20 muestra el cruce entre género y asistencia al preescolar. Como se observa, la proporción de varones que asistieron al preescolar (100/120 = 83 %) es mayor que la proporción de mujeres que lo hicieron (80/111 = 72 %). ¿Qué tan significativa puede ser esta diferencia? ¿Existe una asociación entre las dos variables?

Tabla 20. Cruce entre las variables de género y asistencia al preescolar

		Género		Total
		Masculino	Femenino	
Estudió preescolar	Sí	100	80	180
	No	20	31	51
Total		120	111	231

Los resultados sobre las medidas de asociación se muestran en la tabla 21 tal y como los presenta el IBM-SPSS. Para publicaciones científicas, en general, se recomienda presentar este tipo de resultados en el texto y no repetir los mismos resultados en diferentes formatos (APA, 2010).

Tabla 21. Salida del SPSS sobre medidas de asociación para variables nominales

		Valor	Significación aproximada
Nominal por nominal	Phi	,136	,039
	V de Cramer	,136	,039
	Coeficiente de contingencia	,134	,039
N de casos válidos		231	

Una forma de presentar estos resultados en el texto podría ser como sigue:

> *El cálculo de las medidas de asociación entre estas dos variables indica la presencia de asociaciones significativas para las dos medidas consideradas* Φ *=,136* p*=,039. Aunque la asociación es más bien pequeña, es significativa y podría indicar una situación de inequidad de género: los varones presentes en la muestra asistieron a la educación preescolar en mayor proporción que sus compañeras.*

En un segundo ejercicio, se examinará la asociación entre género y haber estudiado en otros colegios. Se omite en este caso la tabla con los resultados de las medidas de asociación y sus niveles de significación. Los resultados podrían ser presentados como en la tabla 22. Los resultados del cruce entre el género y el haber cambiado de colegio se presentan en la tabla. Aunque existen diferencias, estas parecen ser muy pequeñas, y así lo atestiguan las medidas de asociación examinadas: no se presentan asociaciones significativas entre las dos variables Φ=-,016 *p*=,803.

Tabla 22. Tabla de contingencia entre género y haber estudiado en otros colegios

		Género		Total
		Masculino	Femenino	
¿Ha estudiado en otros colegios?	Sí	88	83	171
	No	32	28	60
Total		120	111	231

Capítulo 5

Regresión lineal

Presentación

En el presente capítulo se trabajarán los temas de regresión lineal simple y múltiple, como parte de la estadística descriptiva. Aunque esto se hace como una forma de generalizar la correlación, al usar paquetes estadísticos como el SPSS, las salidas muestran pruebas que ya son parte de los temas de la estadística inferencial, tales como la prueba *t* de Student o el análisis de varianza (Anova). Como lo hicimos en el capítulo de correlación, explicaremos lo relacionado con el uso de estas pruebas en las regresiones de forma superficial, y pospondremos un tratamiento más detallado para el momento que abordemos propiamente el estudio específico de esas pruebas en la parte dedicada a la estadística inferencial.

Regresión simple y correlación

En esta sección estudiaremos la regresión lineal simple; esto es, la construcción de una recta que describe una relación lineal entre dos variables. Más adelante generalizaremos este concepto a la regresión lineal múltiple, en la que examinaremos la construcción de una recta con múltiples variables predictoras.

La regresión lineal simple y la correlación están estrechamente relacionadas. En principio, ambas parten de dos variables cuyos valores están determinados en el mismo conjunto de población. La correlación tiene que ver con el sentido y la magnitud de la relación entre las dos variables. La regresión utiliza esa relación para poder *predecir* el valor de una variable, conociendo el valor de la otra.

Para poder hacerlo, ahora tendremos que empezar a diferenciar el papel de cada una de las dos variables involucradas. Por un lado, tendremos que identificar la *variable predictora*, también conocida como *variable independiente*, a partir de la cual podremos determinar los valores de la variable predicha, más conocida como *variable dependiente*, o *variable de criterio.*

Esto nos representa un cambio importante frente a la forma en que conceptualizamos la relación entre las dos variables. En términos de la correlación, tal y como las estudiamos en el capítulo anterior, sabemos que la correlación es simétrica; es lo mismo considerar la correlación de X y Y que la de Y y X. No hay diferencia. En términos de la regresión sí hay diferencia. Si X y Y están correlacionadas, un problema será determinar los valores de Y sabiendo los valores de X; otro, diferente,

será determinar los valores de X conociendo los valores de Y. Iniciemos, entonces, considerando que queremos predecir los valores de Y, conociendo los de X.

La construcción de la recta de regresión: predecir Y con X

Aspectos conceptuales

Ya en el capítulo anterior habíamos considerado que, cuando dos variables están correlacionadas, es posible que pudiéramos, conociendo el valor de una, anticipar el valor que tendría la otra. Cuando la correlación es perfecta, esto no debe ofrecer mayor problema, ya que existe una recta que pasa por todos los puntos. El asunto sería más de tipo técnico: determinar cómo es esa recta (que tendrá la forma y = mx +b, como todas las rectas) y sustituir los valores de X para obtener los de Y.

Pero, ¿qué hacemos cuando la correlación es imperfecta? Consideremos, por ejemplo, el caso de la correlación más imperfecta posible: cuando r=0. ¿Cuál será la predicción que se puede hacer? Parece evidente que, si no hay correlación, la mejor predicción para cualquier valor de X será la media de la segunda variable (Y). No es una muy buena predicción, pero en este caso, es la mejor posible.

El tema que nos interesa es hacer una buena predicción cuando la correlación entre dos variables está presente, pero es imperfecta. Consideremos el siguiente ejemplo, en el que tenemos información (ficticia) de 25 estudiantes sobre dos variables: 1) el número de horas invertidas durante la semana en estudiar para un examen y 2) la nota obtenida en ese examen. La figura 32 muestra el diagrama de dispersión. Como se observa, la relación es imperfecta, positiva y lineal. El problema de la regresión es encontrar una recta, y solo una, que describa esta relación de la mejor forma posible.

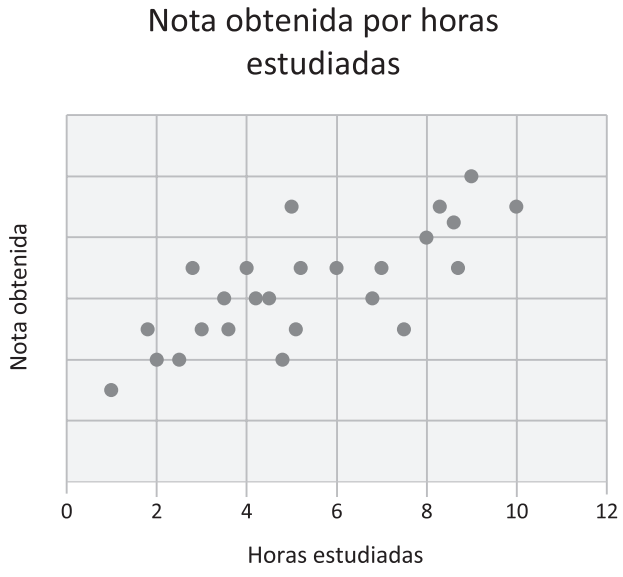


Figura 32. Diagrama de dispersión

La solución a este problema consiste en construir una recta que minimice los errores de predicción, de acuerdo con el llamado *criterio de los mínimos cuadrados*. Explicaremos brevemente en qué consiste este criterio. Supongamos que (X, Y) son las coordenadas de uno de nuestros puntos, y que Y' es el valor de la predicción para X. En ese caso, la diferencia entre Y (valor obtenido) y Y' (valor predicho), esto es,

$$Y - Y'$$

sería el *error* obtenido en esa predicción específica (figura 33).

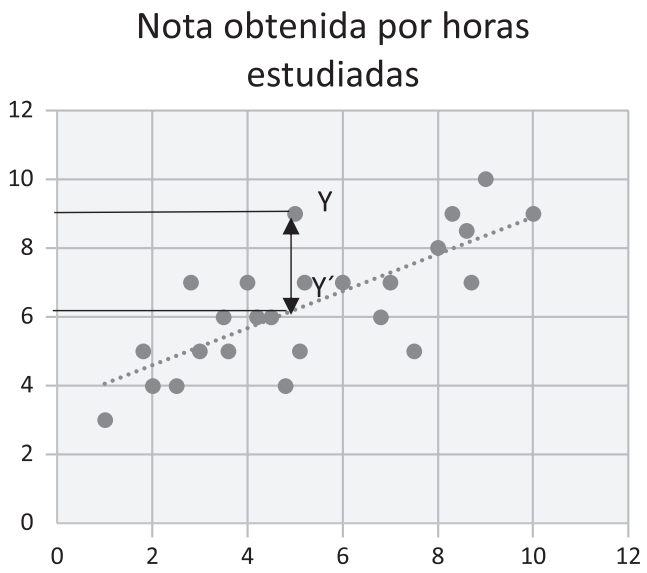


Figura 33. Diagrama de dispersión en el que se ha graficado la recta de regresión y el error (Y-Y') de una predicción específica (Y)

Ahora, si quisiéramos calcular el error total de la predicción, la mejor posibilidad sería sumar todos los errores en todos los puntos. El problema en ese caso es que el valor Y – Y' a veces es positivo, cuando el dato real es mayor que su predicción, y a veces es negativo, cuando pasa lo contrario. Así, si sumáramos todos los errores de predicción (Y–Y´), los errores se cancelarían entre sí.

Un truco que los estadísticos usan con bastante frecuencia, en estos casos, es elevar al cuadrado los errores de predicción de cada punto y hacer, ahora sí, la suma total de los cuadrados de los errores de predicción. Esto se hace porque el cuadrado de cualquier valor real siempre es positivo. Así la *suma de los cuadrados de los errores* es el valor que el procedimiento minimiza para la construcción de la recta de regresión (figura 34).

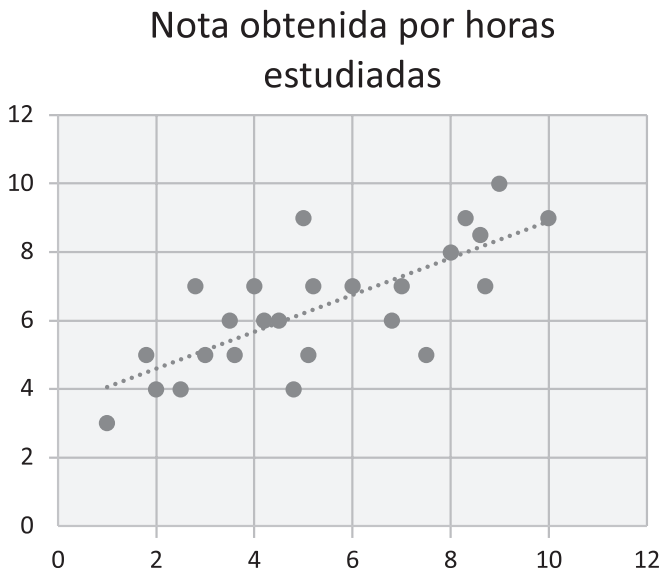


Figura 34. Recta de regresión en el diagrama de dispersión

Como en otras ocasiones, omitiremos las fórmulas específicas. Baste mencionar que la recta de regresión tendrá la siguiente forma:

$$Y' = B_0 + B_1X$$

Donde

Y' es el valor predicho para Y en el punto X
B_1 es la *pendiente* de la recta de regresión, Y
B_0 es el punto del eje en el que la recta cortaría el eje Y o la *constante de regresión*.

Cómo obtener la ecuación en SPSS e interpretar las salidas

Para correr un análisis de regresión en el SPSS, se debe seguir la ruta /Analizar/Regresión/Lineales… Existen muchas posibilidades en esa ventana, pero en este momento, nos limitaremos a especificar cuál es nuestra variable de criterio, o variable dependiente (nota obtenida) y cuál es nuestra variable predictora, o variable independiente (horas estudiadas). Las variables independientes se incluyen en el espacio "Bloque 1 de 1" y, en este momento, en el que estudiamos la regresión simple, la limitaremos a una sola variable. Más adelante, en este capítulo, examinaremos la regresión con más de una variable, que llamamos "regresión múltiple".

Todas las opciones se dejan intactas en el valor predeterminado. Más adelante, cuando estudiemos con mayor profundidad la regresión, modificaremos algunas opciones.

Los resultados en el SPSS son como se muestra en la tabla 23. Primero, se aportan datos sobre la bondad de ajuste del modelo en una tabla titulada "Resumen del modelo".

Tabla 23. Indicadores de bondad de ajuste del modelo de regresión simple

Resumen del modelo				
Modelo	**R**	**R cuadrado**	**R cuadrado ajustado**	**Error estándar de la estimación**
1	,744[a]	,554	,534	1,25848

a. Predictores: (Constante), V1 Horas estudiadas

El primer valor es titulado R. En este caso, en el que estamos construyendo una regresión simple, "R" coincide con nuestro —ya bien conocido— coeficiente de correlación producto-momento de Pearson; esto es, *r*. En nuestro caso R=*r*=,744, que representa una correlación alta.

El valor de "R cuadrado", que también lo habíamos explicado en el capítulo de correlación, se refiere al coeficiente de determinación, o proporción de varianza total explicada. Entre mayor sea *r*, mayor es esta proporción y más potencial predictivo tiene el modelo. En nuestro caso, R^2=,554, que representa que el 55,4 % de la varianza total de la variable dependiente puede ser explicada por el modelo. Este valor siempre debe ser reportado.

La tercera columna se titula "R cuadrado ajustado". Esto representa una corrección al valor de r^2 que se basa en el número de casos y en el número de variables independientes. En una situación con pocos casos y muchas variables independientes r^2 puede ser artificialmente alto, y entonces el valor de "R cuadrado ajustado" será mucho más bajo. En nuestro caso, hay pocas variables independientes, solo una, pero también muy pocas observaciones: solo 25. Aun así, el valor de r^2 corregido no es mucho menor que el del r^2 original: solo 0,020 por debajo, que representa apenas un 2 % de pérdida de la proporción de varianza explicada.

La última columna, titulada "Error estándar de la estimación" representa lo contrario a los indicadores anteriores, esto es, la parte de la varianza de la variable dependiente que *no es explicada* por la recta de regresión. Cuanto menor es ese valor, mejor es el ajuste del modelo. Este valor rara vez se reporta.

La segunda salida del SPSS para nuestra regresión es una tabla titulada Anova. El análisis de varianza, comúnmente resumido como Anova, es una prueba estadística diseñada para evaluar si hay diferencias significativas entre los valores de dos o más medias. En este caso, el Anova se utiliza para comparar si hay diferencias entre el potencial predictivo del modelo y el logrado solo por efecto del azar.

Tabla 24. Tabla del análisis de varianza (Anova) en regresiones

Anova[a]						
Modelo		**Suma de cuadrados**	**gl**	**Media cuadrática**	**F**	**Sig.**
1	**Regresión**	45,213	1	45,213	28,548	,000[b]
	Residuo	36,427	23	1,584		
	Total	81,640	24			

a. Variable dependiente: V2 Nota.
b. Predictores: (constante), V1 horas estudiadas.

Para nuestro caso, el valor del F obtenido (28,548) y en particular el de sus niveles de significación (<,001) indican que es extremadamente improbable que hubiéramos obtenido este nivel de *r* (,744) por efecto del azar y que, por lo tanto, es extremadamente probable que entre estas dos variables exista una relación lineal. La interpretación de este tipo de pruebas se desarrollará en detalle más adelante en un capítulo dedicado al tema.

Finalmente, la tercera salida del SPSS nos da la información que hemos estado buscando: la recta de regresión (tabla 25).

Tabla 25. Tabla de coeficientes de regresión

Coeficientes[a]						
Modelo		**Coeficientes no estandarizados**		**Coeficientes estandarizados**	***t***	**Sig.**
		B	**Error estándar**	**Beta**		
1	(Constante)	3,519	,592		5,946	,000
	Horas estudiadas	,538	,101	,744	5,343	,000

a. Variable dependiente: V2 Nota.

Las filas representan, en su orden, la constante de regresión de la recta (constante) y la pendiente de la recta construida (horas estudiadas). Sobre cada una se aportan los siguientes datos:

En la fila de "coeficientes no estandarizados":

- Los B indican los coeficientes de la ecuación de regresión, usados para el cálculo de las puntuaciones directas. Específicamente

  Nota = 3,519 + 0,538 * horas estudiadas

- A cada valor de B, se aporta su error estándar. Este dato, que rara vez se reporta, se usa en el cálculo de los valores *t*, como veremos más adelante.

La columna de los "coeficientes estandarizados beta" (β), coincide, en la regresión simple, con el coeficiente de correlación *r* de Pearson y representa la pendiente de la recta presente entre los valores *z* de las dos variables. En la regresión múltiple, los β permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación de regresión.

Por último, aparecen los valores *t* con sus correspondientes niveles de significación. La prueba *t* de Student, que será estudiada más adelante, permite comparar medias y establecer los niveles de significación de la diferencia. En este caso, nos permiten contrastar la hipótesis de que los coeficientes de regresión valgan cero para toda la población.

En nuestro caso particular, los resultados de estas últimas dos columnas nos permiten afirmar que

- La constante de la recta de regresión difiere significativamente de 0. Este dato, en general, no es importante.
- La pendiente de la recta difiere significativamente de 0, lo que nos permite concluir que hay una relación lineal estadísticamente significativa (p<,001) entre las dos variables. Recuérdese la significación estadística que examinamos en las correlaciones lineales en el capítulo anterior.

Aspectos importantes para la interpretación de regresiones

En este momento debemos tener en cuenta dos puntos fundamentales para la interpretación de regresiones. El primero, que ya lo habíamos mencionado para las correlaciones, es evitar hacer inferencias sobre relaciones de causalidad. El segundo es tener precaución con valores por fuera del rango observado de las variables.

Regresión no es causalidad

Al respecto del primer punto, y aunque ya lo hemos dicho ampliamente, no sobra reiterar en este espacio que una cosa es poder predecir los valores de una variable conociendo otra, y otra muy diferente es suponer que haya una relación de causalidad entre las dos variables. La regresión toma base en la correlación, y la correlación, como ya hemos dicho, es solo covariación, no causalidad. Así, por mucho que nos sintamos tentados a ver en los resultados de una regresión una demostración de una relación causal, es imperativo no hacerlo. Podemos predecir sin suponer causalidad.

Extrapolación a valores por fuera del rango de la muestra

El segundo punto es más técnico. En general, calculamos regresiones para predecir el valor de una variable (de criterio) por los valores de otra variable, u otras, en poblaciones que no hacen parte de nuestra muestra. Esto es obvio ¿para qué podríamos querer predecir datos que ya tenemos? Sin embargo, esto también nos pone en un problema diferente, puesto que los datos de la población pueden superar los parámetros de los datos presentes en la muestra misma.

En nuestro ejemplo, podríamos querer extrapolar nuestros datos para el caso de una persona que haya estudiado 15 horas. Al sustituir ese valor en la ecuación, ese hipotético estudiante obtendría una nota de 11,586, lo cual es imposible. Otro caso, ausente en nuestra muestra, es el de un estudiante que no haya estudiado en absoluto; esto es, que ha estudiado 0 horas. La sustitución de ese valor en la ecuación nos indica que este estudiante obtendrá una nota de 3,519, pero la verdad es que dentro de nuestra muestra no hubo ningún estudiante que hubiera estudiado 0 horas, por lo que no sabemos si, en ese rango, la relación seguiría siendo lineal.

Regresión de X sobre Y

En la sección anterior examinamos la regresión de la nota obtenida por el número de horas estudiadas. Esta regresión tiene un cierto sentido intuitivo en la medida en que entendemos que la nota obtenida es, al menos en parte, la consecuencia del número de horas estudiadas. Tenemos, en este caso, una idea de que una variable es causa de la otra y que podemos utilizar el conocimiento que tenemos sobre la causa y la relación que sabemos que existe para hacer predicciones sobre la consecuencia.

Un punto que es muy importante en relación con las regresiones, así como lo fue con las correlaciones, es que predecir no significa, en ningún caso, asumir una relación de causalidad entre las variables. No necesitamos asumir causalidad para hacer predicciones; dicho de otra forma, poder hacer buenas predicciones no valida, de ninguna manera, una relación de causalidad entre las variables.

En realidad, y a pesar de tener esta relativa claridad sobre las relaciones entre las variables, nada nos impide ahora considerar el establecimiento de la regresión en el sentido inverso; esto es, conociendo la nota obtenida, intentar hacer una "predicción" sobre el número de horas estudiadas. El diagrama de dispersión invierte las variables presentadas en la figura 35.

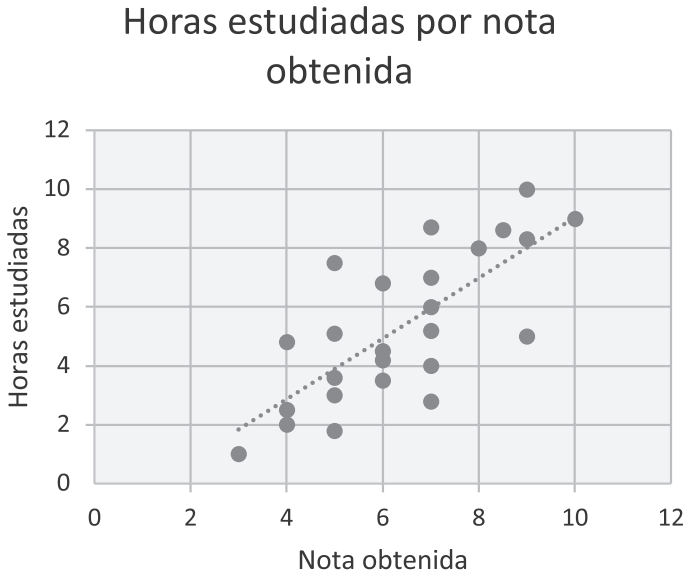


Figura 35. Recta de regresión para la predicción de las horas estudiadas conociendo la nota obtenida

Nuestro interés en esta sección es calcular la nueva recta de regresión y examinar cuáles son los elementos constantes y cuáles varían entre las dos rectas.

Primero, examinemos los indicadores de bondad de ajuste del nuevo modelo. Los datos se presentan en la tabla 26. Tal y como se observa, nuestro nuevo modelo lineal mantiene los mismos valores de R, R cuadrado y R cuadrado ajustado. Esto es apenas natural por cuanto, como se recordará, el coeficiente de correlación de Pearson (r) entre las dos variables es el mismo, por lo que sus valores r^2 y r^2 ajustado permanecen iguales. El error estándar (EE) de la estimación, sin embargo, es diferente en los dos modelos: en el primer caso EEyx=1,25848, y ahora EExy=1,73993. Esto ocurre por las diferencias en los valores de la desviación estándar de las dos variables.

Tabla 26. Indicadores de bondad de ajuste del modelo que predice las horas estudiadas

Resumen del modelo				
Modelo	**R**	**R cuadrado**	**R cuadrado ajustado**	**Error estándar de la estimación**
1	,744[a]	,554	,534	1,73993

a. Predictores: (constante), nota.

En la segunda salida, se examina la tabla de Anova que analiza la hipótesis de que el modelo es significativamente mejor que el azar. Los resultados arrojados por el SPSS se presentan en la tabla 27.

Tabla 27. Análisis de varianza del modelo de número de horas estudiadas

Anova[a]						
Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	86,424	1	86,424	28,548	,000[b]
	Residuo	69,629	23	3,027		
	Total	156,054	24			

a. Variable dependiente: horas estudiadas;

b. Predictores: (constante), nota.

Una comparación entre los datos presentados y los previamente obtenidos muestra que, salvo los de suma de cuadrados y media cuadrática, los valores fundamentales del Anova permanecen idénticos entre los dos modelos, y muy particularmente los relacionados con el valor de F, su nivel de significación y los grados de libertad (gl).

Por último, debemos examinar los datos relacionados con los coeficientes de regresión. Los resultados se presentan en la tabla 28.

Tabla 28. Coeficientes en la ecuación de regresión del número de horas estudiadas

Coeficientes[a]						
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	*t*	Sig.
		B	Error estándar	Beta		
1	(Constante)	-1,248	1,277		-,978	,338
	Nota	1,029	,193	,744	5,343	,000

a. Variable dependiente: horas estudiadas.

Tal y como se observa, los coeficientes no estandarizados de regresión difieren entre los dos modelos. Esto es natural y era completamente esperable. En el nuevo modelo la constante B_0=-1,248 y la prueba *t* muestra que este valor no difiere significativamente de cero *t*=-,938 *p*=,338, que era algo que no ocurría en el modelo anterior. Por su parte, la pendiente de la recta es de B_1=1,029, bastante mayor que la obtenida en el modelo anterior (0,538), lo que indica que nuestra nueva recta está más "empinada" que la anterior. Así, nuestra nueva recta de regresión es:

Horas estudiadas = -1,248 + 1,029 * nota

Por otro lado, y como ya lo esperábamos, el valor del coeficiente estandarizado (β) para la pendiente, obtenido para este modelo, es idéntico al obtenido en el modelo anterior, ya que coincide con el coeficiente de correlación de Pearson entre las dos variables. Más aún: el valor obtenido por la prueba *t*, que mide la importancia relativa de β, es idéntico al obtenido en el caso anterior *t*=5,343 *p*<,001 y, por supuesto, indica lo mismo: que la pendiente de la recta difiere significativamente de cero, por lo que debe asumirse una relación lineal entre las variables.

Ejemplo (datos reales)

En una de las bases que hemos venido trabajando tenemos información acerca de la aplicación de dos pruebas estandarizadas a una muestra grande de la población estudiantil de secundaria en el grado 10 (N=1499): estas son las pruebas de Matemáticas y Ciencias. En particular, examinaremos ¿en qué medida podemos predecir el puntaje en la prueba de Ciencias, si conocemos el puntaje en las pruebas de Matemáticas? La figura 36 muestra el diagrama de dispersión entre estas dos variables. Como se observa, parece verificarse una relación lineal.

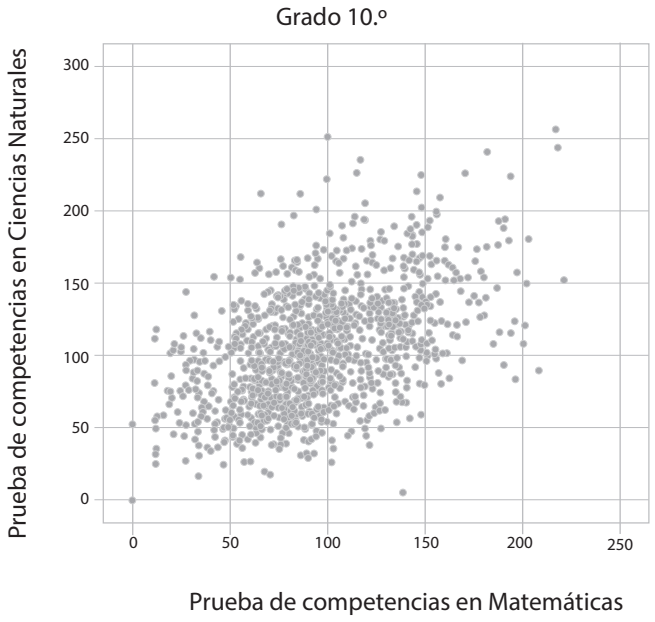


Figura 36. Diagrama de dispersión de las variables "Prueba de competencias en Ciencias Naturales" y "Prueba de competencias en Matemáticas"

La tabla 29 muestra los indicadores de bondad de ajuste del modelo. Tal y como se observa, tenemos un coeficiente de correlación producto momento de Pearson de *r*=,488. Correspondientemente, el coeficiente de determinación alcanza el valor de ,238, y el R^2 ajustado presenta ese mismo valor, en gracia al tamaño de la muestra, lo que indica que podemos explicar el 23,8 % de la variación del puntaje en Ciencias por el puntaje en Matemáticas. La prueba del análisis de varianza indica que el modelo lineal predice de forma muy significativa la variable dependiente.

Tabla 29. Indicadores de bondad de ajuste

Resumen del modelo[a]				
Modelo	**R**	**R cuadrado**	**R cuadrado ajustado**	**Error estándar de la estimación**
1	,488[b]	,238	,238	33,406

a. Grado = Grado 10.°

b. Predictores: (constante), pba_mat prueba de competencias en Matemáticas.

Anova[a,b]						
Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	433032,926	1	433032,926	388,025	,000[c]
	Residuo	1383831,545	1240	1115,993		
	Total	1816864,471	1241			

a. Grado = 2 Grado 10.°.

b. Variable dependiente: prueba de competencias en Ciencias Naturales.

c. Predictores: (constante), prueba de competencias en Matemáticas.

Por último, la tabla 30 muestra los coeficientes de regresión. La ecuación de regresión es:

PCiencias= 59,343 + 0,483*PMatemáticas

El análisis del valor (beta) indica que la relación lineal entre estas variables es ampliamente significativa.

Tabla 30. Coeficientes del modelo de regresión lineal simple del puntaje de Ciencias

Coeficientes[a,b]						
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	*t*	Sig.
		B	Error estándar	Beta		
1	(Constante)	59,343	2,481		23,920	,000
	Prueba de competencias en Matemáticas	,483	,025	,488	19,698	,000

a. Grado = Grado 10.°.

b. Variable dependiente: Prueba de competencias en Ciencias Naturales.

Regresión lineal múltiple

La regresión lineal múltiple es, sencillamente, la extensión de la regresión lineal simple al caso en el que tenemos una variable dependiente (predicha), y dos o más variables predictoras.

La forma general de la ecuación de regresión para *n* variables predictoras (X_1, X_2… X_n) es

$$Y' = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_nX_n$$

Como se observa, esta ecuación es muy similar a la de regresión simple. Solo hemos añadido las otras variables involucradas (X_i) con sus respectivos coeficientes (B_i).

Para el examen de la forma en que se interpretan las salidas del SPSS en una regresión múltiple extenderemos uno de los ejemplos que examinamos en las regresiones simples. En ese caso examinamos cómo podíamos predecir el puntaje de la prueba de Ciencias conociendo el puntaje de la

prueba de Matemáticas. Ahora extenderemos el análisis para incluir, en este modelo, el puntaje de la prueba de Lenguaje. Así, ¿en qué medida podemos predecir el puntaje en la prueba de Ciencias, si conocemos el puntaje en las pruebas de Matemáticas y Lenguaje?

Los resultados se presentan en las tablas 31 y 32. En este caso, R ya no representa un coeficiente de correlación entre dos variables sino el *coeficiente de correlación múltiple* entre las tres variables involucradas. Esta es una extensión del concepto de correlación lineal que ya vimos. Puede ser interesante notar que este R es siempre mayor que la mayor correlación presente entre la variable dependiente y las independientes. En nuestro caso, el coeficiente de correlación múltiple es R=,561; la correlación entre la prueba de Ciencias y la de Matemáticas es de *r*=,488 y la presente entre la prueba de Ciencias y la de Lenguaje es *r*=,457.

Correspondientemente, ahora el R cuadrado representa el *coeficiente de determinación múltiple*, que resulta ser una extensión del coeficiente de determinación que mantiene el mismo significado: la proporción de la varianza del puntaje de Ciencias que explicada por el modelo. El R cuadrado ajustado apenas difiere del R cuadrado, debido al gran tamaño de la muestra. En ese sentido y, tal y como se observa, tenemos un buen modelo, que alcanza a explicar hasta el 31,5 % de la varianza de la prueba de Ciencias (R^2). Recuérdese que cuando solo incluíamos el puntaje de Matemáticas en el modelo el coeficiente de determinación llegaba apenas hasta R^2=,238. Esto significa que al incluir la nueva variable hemos explicado un 7,7 % adicional de la varianza de la variable dependiente.

Tabla 31. Indicadores de bondad de ajuste del modelo de regresión lineal múltiple

Resumen del modelo[a]				
Modelo	**R**	**R cuadrado**	**R cuadrado ajustado**	**Error estándar de la estimación**
1	,561[b]	,315	,314	31,691

a. Grado = grado 10.°.

b. Predictores: (constante), pruebas de competencias en Lenguaje, prueba de competencias en Matemáticas.

Tabla 32. Análisis de varianza del modelo de regresión lineal múltiple

Anova[a,b]						
Modelo		**Suma de cuadrados**	**gl**	**Media cuadrática**	**F**	**Sig.**
1	**Regresión**	572500,263	2	286250,132	285,016	,000[c]
	Residuo	1244364,208	1239	1004,329		
	Total	1816864,471	1241			

a. Grado = grado 10.°.

b. Variable dependiente: prueba de competencias en Ciencias Naturales.

c. Predictores: (constante), pruebas de competencias en Lenguaje, prueba de competencias en Matemáticas.

En la tabla 32 se observa que el modelo predice mejor el valor de la variable dependiente que el azar de forma estadísticamente significativa. Finalmente, la tabla 33, de coeficientes, nos indica los valores específicos de la ecuación de regresión. Específicamente, sabemos ahora que

$$PCiencias= 37{,}247 + 0{,}356*PMatemáticas + 0{,}258*PLenguaje$$

Otro punto muy importante es que ahora sabemos que tanto la prueba de Matemáticas como la de Lenguaje son predictores muy significativos de la prueba de Ciencias. Aún más, sabemos que la prueba de Matemáticas es un predictor más fuerte que la del Lenguaje, en la medida en que el valor β de la prueba de Matemáticas (0,359) es mayor que el β de la prueba de Lenguaje (0,306).

Tabla 33. Tabla de coeficientes

Coeficientes[a,b]						
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	37,247	3,009		12,378	,000
	Prueba de competencias en Matemáticas	,356	,026	,359	13,857	,000
	Pruebas de competencias en Lenguaje	,258	,022	,306	11,784	,000

a. Grado = grado 10.°.

b. Variable dependiente: prueba de competencias en Ciencias Naturales.

La regresión múltiple presenta algunas complejidades adicionales al caso de la regresión simple, especialmente relacionadas con la presencia de correlaciones entre las variables independientes. En efecto, si bien ahora cada una de las correlaciones entre la variable dependiente y las variables independientes tendrán un cierto poder predictivo, al considerarse las correlaciones entre las variables independientes, notaremos que parte del poder predictivo de una variable resulta redundante al intersecarse con aquello que también es parte del poder predictivo de otra variable. Por esta razón, los valores β de las variables en la regresión múltiple son menores que la correlación entre la variable dependiente y la independiente. En este tipo de regresión, los valores β se calculan de forma que represente la contribución única de la variable, excluyendo cualquier superposición con las otras variables en la ecuación.

Capítulo 6

Validez, confiabilidad, análisis de escalas y análisis de ítems

Con mucha frecuencia, los investigadores sociales y educativos utilizan diferentes tipos de cuestionarios para indicar algunas de las variables específicas que constituyen sus objetos de estudio. A veces, la investigación misma requiere del desarrollo de un nuevo cuestionario. Este tipo de investigación, de naturaleza metodológica, demanda el uso profundo de una serie de conocimientos que pueden ser englobados en la rama de la psicología que trata de los temas de la medición psicológica: la *psicometría*. La exposición detallada de ese cuerpo de conocimiento, de naturaleza altamente especializada excede objeto de nuestro trabajo. Recomendamos al lector interesado en una introducción al tema, la lectura del texto clásico de Anastasi y Urbina (1998), o bien el abordaje de los trabajos de J. Muñiz (1992), si bien estos últimos pueden requerir de mayor formación matemática.

Por otra parte, en la mayoría de las investigaciones que usan cuestionarios, se utilizan algunos que han sido desarrollados previamente y sobre los cuales se conocen sus características técnicas a partir de un historial de aplicaciones. Al respecto, es importante insistir en que los instrumentos utilizados deben validarse previamente en contextos amplios, preferiblemente internacionales, y deben calcularse sus características técnicas. Aun así, se requiere que, en cada nueva aplicación, se examine el funcionamiento del cuestionario en la muestra. Esto significa que es necesario examinar los datos de validez del cuestionario y calcular su confiabilidad. En este capítulo expondremos la forma de hacerlo.

Los conceptos de validez y confiabilidad

Validez

La *validez* de los instrumentos psicológicos se refiere a lo que estos verdaderamente miden. Comúnmente, se dice que la validez de un instrumento es el grado en que el instrumento mide lo que *dice* medir.

El concepto de validez en las pruebas ha tenido una evolución histórica importante. Originalmente, se aplicaba a instrumentos que evaluaban lo aprendido en determinadas áreas, por lo que la validez se refería a la comparación entre el contenido de la prueba y el contenido del área que se pretendía evaluar. Más adelante, el énfasis se hizo en la capacidad de predicción de la prueba.

En este sentido, la validez quedaba indicada por una correlación entre los resultados de la prueba y la medida de criterio. Actualmente, el término pone mayor énfasis en los aspectos teóricos de aquello que se pretende medir y examina en detalle la verificación empírica de las hipótesis que se derivan del constructo teórico (Anastasi y Urbina, 1998).

En general, vale la pena distinguir entre diferentes tipos de validez.

- *Validez de contenido*. Una prueba tiene cierta validez de contenido cuando el tema tratado por sus ítems es representativo de la destreza o del área que se pretende medir. En esta intentamos establecer la medida en que todos los elementos importantes del contenido estén adecuadamente representados en los ítems, de forma proporcional a su importancia relativa dentro del área de conocimiento y que elementos externos al área no ejerzan excesiva influencia. Usualmente, ese tipo de validez se establece mediante el juicio de expertos reconocidos en el área. En estos casos, con frecuencia, se desarrollan instrumentos a para que los responda el grupo de expertos y se examinan sus respuestas.
- *Validez de criterio o predictiva.* En esta se examina la capacidad que presenta un instrumento para predecir la ejecución que tendrá el sujeto en algún área de desempeño. Esta área la denominamos "criterio". Usualmente, se calcula la validez predictiva de un instrumento mediante el cálculo de la correlación entre los resultados del instrumento y el rendimiento del criterio. Este método requiere, como es obvio, de un intervalo de tiempo presente entre las dos medidas.
- *Validez concurrente*. Se refiere a la situación en la que la puntuación del cuestionario y la del criterio se obtienen en el mismo momento de tiempo. En este caso no se está estableciendo una predicción, sino la relación entre dos medidas. Con frecuencia, se utiliza cuando se pretende ratificar una forma más corta de un cuestionario que ha sido previamente validado. En este caso, como en el anterior, se requiere la comparación entre los resultados del instrumento y el criterio dado y, en esa medida, es la relación con el criterio lo que valida el instrumento.
- *Validez de constructo*. Indica la medida en que los resultados de un instrumento se comportan de acuerdo a lo que teóricamente se espera del él. Esto significa que las puntuaciones del instrumento deben correlacionarse con ciertas variables que se espera, desde la teoría, estén relacionadas y, al tiempo, no tienen que relacionarse con otras con las cuales no se esperan correlaciones. Este es el tipo de validez más cercano al concepto general de validez de un instrumento. Uno de los procedimientos más comunes para establecer la validez de constructo de un instrumento consiste en realizar un análisis factorial, exploratorio o confirmatorio, de los ítems del instrumento. El resultado de este análisis debe coincidir con lo que la teoría indica sobre las diferentes dimensiones presentes en el constructo y las relaciones entre ellas. Este tipo de validez también ha sido denominada "validez factorial".
- *Validez aparente, o validez de apariencia.* Para este tipo de validez se requiere que los ítems de un instrumento tengan la apariencia, a ojos del sujeto evaluado, de evaluar lo que dicen evaluar. Su importancia está dada por el control de un sesgo que puede aparecer en la persona que realiza el cuestionario y que falsea sus ejecuciones (Jensen, 1980).

A menos que la investigación tenga como sus objetivos el desarrollo de instrumentos de medición psicológica, no es frecuente que se examinen los datos de validez de los instrumentos utilizados. Por el contrario, la validez de los instrumentos se entiende garantizada por el historial de su uso. Otro es el caso de examen de la confiabilidad, que se acostumbra a verificar en cada nueva aplicación de un instrumento y que pasaremos a exponer a continuación.

Confiabilidad

En general, se entiende la *confiabilidad* de una prueba como la consistencia de las puntuaciones obtenidas por las mismas personas cuando se les examina en distintas ocasiones con la misma prueba, con conjuntos equivalentes de ítems o en otras condiciones de examinación (Anastasi y Urbina, 1998).

En términos técnicos, la medición de la confiabilidad de una prueba consiste en estimar qué proporciones de la varianza total de las puntuaciones se debe a "verdaderas diferencias" en las características consideradas y qué proporción puede deberse a errores fortuitos.

Tal y como lo anotan Anastasi y Urbina (1998), existen tantas variedades de la confiabilidad como condiciones que afecten los resultados de una prueba. No se trata de exponer un texto detallado sobre este concepto, sino, siguiendo el espíritu de esta obra, de facilitar al investigador novel que utiliza un instrumento psicológico, el cálculo y el reporte de una medida adecuada de confiabilidad de su cuestionario. Aun así, vale la pena exponer brevemente algunas de las formas de cálculo de la confiabilidad más comunes. Nos detendremos, al final de la exposición, en los métodos más encontrados.

- *Confiabilidad test-retest.* Es el método más obvio de cálculo. Se trata de aplicar el mismo instrumento en dos ocasiones, separadas por un intervalo de tiempo, y examinar la correlación entre las dos puntuaciones. Esta técnica es simple y directa, pero presenta dificultades a la hora de determinar la longitud del intervalo entre las dos aplicaciones y los efectos de la memoria y el aprendizaje.
- *Confiabilidad de formas alternas.* Equivalente a la anterior, pero con el uso de una forma alterna del mismo cuestionario. El uso de formas alternas del mismo instrumento es una manera de evitar las dificultades de la confiabilidad test-retest. Por supuesto, las dificultades de esta técnica se relacionan con la posibilidad de garantizar que las dos formas son, verdaderamente, "formas alternas".
- *Confiabilidad de la división por mitades.* Existen diferentes procedimientos para el cálculo de la confiabilidad con una sola aplicación de la prueba. Como es obvio, esta técnica proporciona una medida de la consistencia del contenido de la prueba y no de estabilidad temporal de las mediciones. En general, se recomienda separar los ítems pares e impares y calcular la correlación entre las dos mitades.
- *Confiabilidad de Kuder-Richardson* y *coeficiente alfa (α) de Cronbach.* Este es el más popular de los métodos para el cálculo de la confiabilidad. Como en el caso anterior, se basa en una única aplicación del instrumento y, por tanto, trata del nivel de consistencia entre ítems. La forma más común de cálculo se conoce como la "fórmula 20 de Kuder-Richardson" (KR20)

(Kuder y Richardson, 1937), que, según se demostró después, se corresponde con la media de todos los coeficientes de correlación que resultan de las diferentes divisiones de una prueba (Cronbach, 1951). Una generalización de esta fórmula, para el caso de ítems expresados en escalas ordinales, se conoce como *el coeficiente alfa (α) de Cronbach*. El alfa de Cronbach es, sin lugar a dudas, la medida de confiabilidad más frecuentemente encontrada en publicaciones científicas en gracia de su simplicidad y la facilidad que tiene su cálculo. En general, aceptamos como válidos los coeficientes alfa mayores o iguales a 0,70, aunque en algunos casos este límite puede bajar hasta 0,65.

- *Coeficiente omega (ω) de McDonald.* A pesar de la popularidad del alfa de Cronbach, algunos autores han mencionado que este coeficiente presenta algunos problemas importantes, ya que el coeficiente se encuentra directamente relacionado con la cantidad de ítems en una prueba, el número de alternativas de respuesta y la proporción de varianza del test (Domínguez-Lara y Merino-Soto, 2015). Por estas razones, además de su mayor sensibilidad frente a otros estimadores (Zinbarg *et al.*, 2005), se ha preferido, progresivamente, el uso de una medida similar al alfa de Cronbach que no presenta estas dificultades: el *valor omega (ω) de McDonald***.** Este coeficiente se obtiene de las cargas factoriales de los ítems en un análisis factorial confirmatorio (AFC), si bien en ocasiones se debe recurrir a otro tipo de análisis factorial, conocido como análisis de factores principales. La interpretación del coeficiente omega, y de los valores considerados como aceptables, es idéntica a la del alfa de Cronbach. Actualmente, la gran mayoría de los paquetes estadísticos permiten el cálculo tanto del alfa como del omega, si bien ya algunos tienen a esta última medida como su procedimiento por defecto (JASP, por ejemplo).
- *Confiabilidad entre calificadores.* En algunos instrumentos una fuente importante de la varianza del error corresponde al juicio subjetivo entre calificadores. Esto es especialmente cierto en pruebas proyectivas de personalidad o en instrumentos de desempeño en una conducta compleja, cuya calificación se basa en una rejilla con un uso que queda a juicio del calificador. En estos casos, se solicita a varios calificadores la valoración de cada desempeño y se calcula una medida de confiabilidad entre calificadores. La más común se conoce como el *coeficiente kappa (κ) de Cohen*. Kappa mide el grado de concordancia de evaluaciones nominales u ordinales realizadas por múltiples evaluadores cuando han evaluado las mismas muestras. Un valor de 1 significa un acuerdo perfecto y un valor de 0 representa una situación en que el acuerdo se hubiera obtenido por azar. Esta es la medida más usada cuando se realizan procesos de validez de contenido por expertos. Tiene este doble uso: puede emplearse para identificar confiabilidad en pruebas con preguntas abiertas, o aquellas guiadas por rúbricas, así como para examinar la validez de contenido al identificar el nivel de acuerdo entre los expertos sobre la pertinencia de un ítem en una escala y sus características.

Recomendaciones

Análisis de consistencia de la escala completa

Los métodos mencionados permiten examinar diferentes aspectos de la confiabilidad de una escala completa. Como se ha mencionado, el modelo más comúnmente utilizado para el examen de la consistencia interna es el del alfa de Cronbach, para el cual se aceptan valores iguales o superiores

a 0,70 (para algunos investigadores, se pueden aceptar valores iguales o superiores a 0,65). Este modelo ha sido progresivamente sustituido por el del omega de McDonald, para el cual se consideran los mismos valores de aceptación.

Análisis de consistencia de los ítems en la escala

En algunas situaciones, el examen de la confiabilidad de la escala completa no es suficiente y se desea examinar el aporte a la confiabilidad de cada uno de los ítems incluidos. Esto es especialmente frecuente cuando los investigadores están desarrollando la escala que se encuentra bajo análisis.

Para el análisis de cada uno de los ítems en una escala se procede de una o dos formas diferentes, en las cuales examinamos las relaciones entre el ítem y el resto de la escala de la que hace parte.

La primera forma de análisis de ítems es muy sencilla y se hace a través de la correlación de Pearson entre el ítem en cuestión y el resto de la escala. Este indicador de consistencia del ítem se conoce como *correlación ítem-total corregida* (CITC). Se trata, básicamente, de examinar la correlación producto-momento de Pearson entre el ítem y la escala completa, pero de la cual se ha excluido el ítem mismo. Esta corrección se hace para evitar el aumento artificial del indicador que resultaría del examen de la correlación entre el ítem y sí mismo dentro de la escala.

Para el examen de la consistencia del ítem con el indicador de correlación ítem-total corregida, se propone la siguiente decisión (Ebel, 1965):

- *CITC > 0,2. Ítem consistente*. En general esperamos que los ítems consistentes tengan correlaciones ítem-total corregidas positivas y superiores a 0,2.
- *-0,2< CITC < 0,2. Ítem no consistente*. Los ítems que muestren CITC menores a 0,2, en valor absoluto, no muestran suficiente consistencia con la escala y bajarán los indicadores de confiabilidad de la escala completa, por lo que debe considerarse su modificación o su eliminación.
- *CITC < -0,2. Ítem con consistencia inversa***.** En el caso en que encontremos CITC negativas y menores a -0,2, tendremos el caso de un ítem que apunta en la dirección claramente contraria a la que apunta el resto de la escala. En este caso, tendríamos la situación de que, si el ítem se expresara de forma invertida, mostraría niveles adecuados de consistencia. Es posible que este ítem haya sido formulado en términos inversos a la escala (ítem inverso), por lo que debe ser recodificado para invertirlo, o bien declarado como ítem inverso —si el programa lo permite—. En cualquier caso, este ítem debe ser cuidadosamente examinado.

La segunda forma para el examen de los ítems individuales es bastante similar, pero en este caso utilizamos la variación del indicador de consistencia interna adoptado, ya sea el alfa o el omega, como resultado de la eliminación del ítem en la escala. En general, se considera que los ítems consistentes no deben aumentar el valor del coeficiente de confiabilidad elegido como resultado de su eliminación. En otras palabras:

- $\omega_{t\text{-}i} \leq \omega_t$. *Ítem consistente*. Si el valor de la confiabilidad disminuye con la eliminación del ítem, debemos considerar que el ítem es consistente con el resto de la escala y tiene que mantenerse.

- $\omega_{t\text{-}i} > \omega_t$. *Ítem inconsistente.* Por el contrario, si el valor de la confiabilidad de la escala sin el ítem aumenta, debe considerarse que el ítem es inconsistente. Puede ser importante examinar la magnitud de la variación. Si la variación es muy pequeña y la escala tiene una cierta tradición y está bien establecida, tal vez no valga la pena proceder a la eliminación del ítem. Por otro lado, si la magnitud de la variación es grande (mayor que 0,05) el ítem debe ser cuidadosamente examinado y, si la escala se encuentra en elaboración, debe considerarse su eliminación.

Análisis de confiabilidad de la escala y los ítems

Para ejecutar el análisis de confiabilidad de una escala, el *software* que utilizamos provee diferentes posibilidades. En términos generales, se recomienda el análisis de confiabilidad de la escala a través del cálculo del valor omega de McDonald y, en el análisis de ítems, el cálculo de la correlación ítem-total corregida.

Para obtener un análisis de confiabilidad en el programa JASP, se debe proceder, en el directorio principal, a través de la opción *Reliability*. Las opciones recomendadas se presentan en el recuadro 18.

Recuadro 18. Cómo obtener un análisis de confiabilidad en JASP

/Reliability/Unidimensional Reliability
En este punto se deben pasar los ítems que componen la escala a la lista "Variables"
Analysis
Scale Statistics
√ Confidence interval 95% (seleccionada por defecto)
√ McDonalds' ω (seleccionada por defecto)
√ Mean √ SD
Individual Item Statistics
√ McDonald's ω (if item dropped)
√ Item-rest correlation
Reversed-Scaled Items
En este punto se deben seleccionar los ítems codificados en la dirección inversa y pasarlos a la lista Reverse-Scaled Items

Debe señalarse que la posibilidad que ofrece el JASP de definir cuáles son los ítems inversos da una gran comodidad para el examen de la confiabilidad. Esta posibilidad no existe en las versiones actuales de IBM-SPSS. Para obtener un análisis de confiabilidad en IBM-SPSS, se procederá a través del menú "Escala" (recuadro 19).

Si existen ítems inversos en la escala, estos deben ser recodificados de forma previa a su inclusión en el análisis de confiabilidad del instrumento.

Recuadro 19. Cómo obtener un análisis de confiabilidad en IBM-SPSS

/Analizar/Escala/Análisis de fiabilidad…
En este punto usted debe seleccionar todos los ítems y pasarlos a la lista "Elementos"
Modelo: Omega ("Alfa" esta seleccionado por defecto)
Estadísticos…
Descriptivos para
√ Elemento
√ Escala
√ Escala si se elimina el elemento
Pulsar el botón "Continuar"
Pulsar el botón "Aceptar"

Ejemplo

En una publicación reciente, se examinaron algunas de las características técnicas de la prueba de autoconcepto AF5 (García y Musitu, 2014) en una muestra amplia de estudiantes del área de la ciudad de Manizales (Hederich *et al.*, 2022). La prueba AF5 evalúa el autoconcepto en una perspectiva multidimensional; se compone de treinta reactivos (ítems) divididos en cinco dimensiones, a saber, académico-laboral, social, emocional, familiar y física, con seis ítems en cada escala. Para el presente ejemplo, expondremos los resultados del examen de confiabilidad de la escala académico-laboral. Las tablas 34 y 35 muestran los resultados según el JASP, traducidos al castellano. Se han solicitado los indicadores del omega de McDonald y los del alfa de Cronbach, a fin de que el lector pueda constatar sus similitudes.

Tabla 34. Estadísticas de confiabilidad

Estimación	ω de McDonald	α de Cronbach
Estimación puntual	0,848	0,842

Tabla 35. Estadísticas de confiabilidad de elementos individuales

	Si el ítem se elimina		
Item	ω de McDonald	α de Cronbach	CITC
AF501	0,835	0,829	0,556
AF506	0,797	0,791	0,755
AF511	0,848	0,838	0,514
AF516	0,834	0,830	0,585
AF521	0,820	0,810	0,670
AF526	0,802	0,798	0,710

Tal y como se observa de la tabla 34, el valor del omega de McDonald alcanza 0,848, que debe ser considerado aceptable. El valor alfa (0,842) es bastante cercano al anterior.

Por su parte, podemos hacer el análisis de cada uno de los ítems que componen la escala a partir de los datos presentados en la tabla 35. Iniciando con la primera columna, se observa que, salvo en el ítem AF511, el valor del omega de McDonald disminuye al eliminar el ítem correspondiente. En el caso del ítem AF511, el valor del omega no cambia al eliminar el ítem. En conclusión: todos los ítems parecen hacer importantes contribuciones a la escala y son consistentes con esta. Este resultado se confirma si examinamos la última de las columnas de la tabla, que muestra que las correlaciones entre cada uno de los ítems y su total corregido son, en todos los casos, mayores a 0,2; de hecho, resultan mayores a 0,5.

Capítulo 7

Introducción a la inferencia estadística

Iniciamos ahora los temas de la estadística inferencial. En la parte anterior, dedicada a la estadística descriptiva, nuestro interés fue tratar de presentar un conjunto de datos de la forma más clara, completa y concreta posible. Ahora iremos más allá de la descripción de un conjunto de datos, para intentar describir poblaciones completas a partir de pequeñas muestras. Este es el campo de la estadística inferencial.

Para abordar estos temas, requerimos, primero, de una visión general del campo de la inferencia estadística. Existen tres tipos de estudios de inferencia estadística con propósitos y métodos diferentes: 1) los estudios para la estimación de parámetros de la población, 2) los estudios de pruebas de hipótesis y 3) los estudios de construcción de modelos. Estos estudios difieren en aspectos fundamentales, tales como los procedimientos para la selección de la muestra y los criterios de validez que presentan. Examinaremos estos tres tipos de estudios en primer lugar.

En lo que queda del presente trabajo, nos dedicaremos a la presentación de las pruebas de hipótesis. Para hacerlo, debemos antes abordar el estudio previo de varios términos centrales. Primero, el de muestra y su relación con el de población; examinaremos allí las formas en las que se pueden diseñar y obtener las muestras en la investigación educativa y social y las formas en que debemos describirlas. Segundo, debemos presentar, de la manera más general y comprensiva posible, el concepto de probabilidad y la terminología que utilizaremos para enunciarla; este resulta fundamental para poder interpretar los resultados de las pruebas de hipótesis. Finalmente, en el tercer punto, presentaremos una cierta curva que resulta fundamental en la estadística: la curva normal. Sobre esta curva calcularemos la probabilidad de haber obtenido un resultado particular. Estos conceptos serán relacionados en un apartado final.

Inferencia y tipos de inferencia estadística

La inferencia estadística comprende todos los métodos y procedimientos para determinar valores y relaciones de características de las unidades de una población objetivo a partir de observaciones realizadas en una o varias muestras de población.

Aunque los métodos de inferencia estadística partieron de un origen común, actualmente se pueden clasificar, de acuerdo con el alcance e intereses del investigador, en tres grandes categorías:

1) la estimación de valores de los parámetros desconocidos en la población; 2) las pruebas de hipótesis y 3) la construcción de modelos estadísticos. La figura 37 ilustra estas tres grandes ramas de la inferencia.

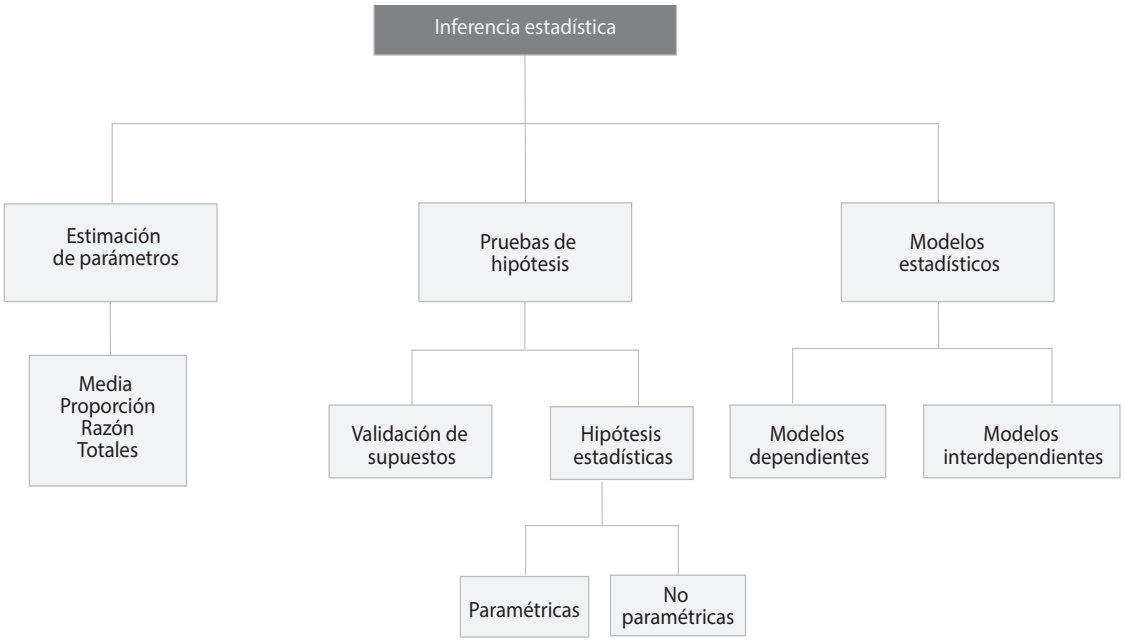


Figura 37. Tipos de Inferencia estadística

La primera rama, dedicada a la estimación de parámetros desconocidos en grandes poblaciones, se fundamenta en la capacidad que tenemos para inferir un parámetro (poblacional) a partir de un estadístico (muestral). El interés en este tipo de estudios es estimar los parámetros poblacionales de la forma más precisa posible, a fin de dar cuenta del estado del objeto estudiado en la población. Ejemplos de este tipo de estudios, realizados en Colombia, o en los cuales el país ha tomado parte, son:

- Estudios sobre prevalencias de consumo de alcohol en la población escolarizada (p. ej. Pérez-Gómez *et al.*, 2018).
- Estudio internacional de educación cívica y ciudadana (Ministerio de Educación Nacional [MEN] e Instituto Colombiano para la Evaluación de la Educación [Icfes], 2017).
- Estudio internacional de educación matemática TIMSS (Mullis *et al.*, 2008a) y estudio internacional de educación en ciencias TIMSS (Mullis *et al.*, 2008).
- Perfiles del docente de instituciones oficiales (p. ej. Londoño *et al.*, 2011)
- Estudios de estilos cognitivos en población estudiantil colombiana (p. ej. Hederich-Martínez y Camargo, 1999; Hederich-Martínez, 2007).
- Estudios de clima y convivencia escolar (Alcaldía Mayor de Bogotá *et al.*, 2016).

En general, en este tipo de estudios se requiere de grandes tamaños de muestra y de estrategias de diseño muestral sofisticadas, tales como muestreo polietápico y estratificado, para optimizar y operacionalizar la muestra. Para el cálculo del tamaño de la muestra, es de fundamental importancia conocer el tamaño de la población y sus posibilidades de estratificación. La calidad de las

estimaciones depende del error estándar de muestreo. Por otro lado, y en gracia a sus grandes tamaños de muestra, en estos estudios es posible profundizar los análisis con métodos multivariados a partir de los datos primarios.

Existen dos formas de estimar parámetros de una población a partir de los estadísticos obtenidos en una muestra: la estimación puntual y la estimación por intervalos. La estimación puntual es clara y sencilla, y consiste en reportar el estadístico muestral. En la estimación por intervalos, por su parte, el dato puntual obtenido en la muestra se presenta dentro de *un intervalo de confianza*, en el cual puede asegurarse que se encuentra el valor real, con un cierto porcentaje de seguridad (usualmente el 95 %, aunque también se reporta el 99 %). El tamaño de estos intervalos es también una medida de calidad de la estimación. Asimismo, esta forma de presentar los estadísticos representará una alternativa a las pruebas de hipótesis, como lo veremos más adelante.

La segunda rama de la inferencia estadística, relacionada con las *pruebas de hipótesis*, también tiene un campo específico en el mundo de la estadística. Su origen se remonta a principios de 1900 y su desarrollo está ligado a la decisión, con pequeñas muestras, en el campo de la experimentación.

En este tipo de estudios el objetivo es construir un estadístico de prueba a partir de una muestra aleatoria que permita evaluar la verosimilitud, en términos de probabilidad, de una aseveración (hipótesis) acerca de una población. En ese sentido, estos estudios no pretenden indagar valores específicos en una población, sino más bien establecer relaciones, asociaciones o diferencias entre variables alrededor de un fenómeno dado. Ejemplos típicos de este tipo de estudio pueden ser:

- Comparar metodologías de enseñanza (p. ej., Vega y Hederich-Martínez, 2015).
- Comparación de métodos de incentivar la lectura en los estudiantes o la escritura en estados iniciales (p. ej., Rincón y Hederich-Martínez, 2012).
- Variables sociodemográficas y académicas asociadas con el "síndrome de *burnout* académico" (agotamiento emocional) en estudiantes universitarios (Caballero *et al.*, 2015).

En estos estudios se requiere un diseño de investigación sofisticado, que posiblemente incluirá diferentes grupos sometidos a diferentes situaciones (experimentales y de control) y diferentes mediciones (antes, durante y después) que puedan ser comparadas. Su desarrollo requiere plantear hipótesis estadísticas (nula y alternativa), seleccionar una muestra aleatoria adecuada, un estadístico de prueba apropiado, calcular su probabilidad bajo la hipótesis nula y tomar decisiones sobre los resultados.

Las pruebas de hipótesis se clasifican en *pruebas paramétricas* y *no paramétricas*, dependiendo de los supuestos subyacentes, los niveles de medición de las variables involucradas y sus distribuciones de probabilidad. Las pruebas paramétricas presentan mayor cantidad de supuestos y de condiciones de aplicación y ofrecen también mayor seguridad en los resultados. Las pruebas no paramétricas tienen menos condiciones y resultan de más fácil aplicación, si bien otorgan también menor seguridad para la decisión.

Los criterios para determinar la calidad de los resultados obtenidos en los estudios de prueba de hipótesis son diferentes a los usados en la estimación de parámetros. Mientras que en la estimación de parámetros era importante el error de muestreo, en este caso son importantes la *potencia de la prueba*, la medición del *tamaño del efecto* y el cálculo del *error estándar de los estimadores*, a

partir del cual se construirán sus *intervalos de confianza*. Todos estos conceptos serán definidos más adelante. En lo que sigue del presente trabajo nos concentraremos en esta rama.

La tercera rama de la inferencia estadística comprende la *construcción de modelos estadísticos*. Esta ha sido la de mayor desarrollo en los últimos años gracias a la disponibilidad de ordenadores y de *software* estadístico. La inferencia en este caso se basa en la construcción de modelos estadísticos a partir de modelos conceptuales. Ejemplos sencillos de estos son los modelos de regresión simple, o múltiple, que ya estudiamos en términos descriptivos en el capítulo 5.

Es posible diferenciar dos tipos de modelos estadísticos: dependientes e interdependientes. En los *modelos dependientes* se asume que existen una o más variables que pueden ser consideradas "de criterio", "dependientes" o de consecuencia de otro grupo de variables que, a su vez, serán consideradas como "variables independientes" y tratamos de predecir las primeras por las segundas. Ejemplos de esto son los diferentes tipos de modelos de regresión (simple, múltiple, logística, ordinal…), análisis discriminante o correlación canónica, por ejemplo. Por su parte, en los *modelos interdependientes* no se asume la presencia de variables dependientes, sino que se examinan las relaciones múltiples entre conjuntos grandes de variables. Ejemplos de modelos interdependientes son el análisis factorial, el análisis de correspondencias múltiples o el análisis de conglomerados (*cluster analisys*), por ejemplo. No profundizaremos, por ahora, en esta rama de la inferencia.

Como se ha mencionado, las formas y procedimientos para el diseño y la selección de la muestra dependen del tipo de estudio y de sus objetivos y, más específicamente, de si se trata de un estudio de estimación de parámetros o de un estudio de prueba de hipótesis. Los diferentes procesos para el diseño y selección de la muestra serán presentados en la siguiente sección.

Población y muestra

La importancia de las muestras

En la investigación educativa y social intentamos construir conocimiento sobre la población, en general. ¿A qué nos referimos con este concepto general de "población"? Esta pregunta no es sencilla. Baste decir que la *población* es un conjunto de sujetos o, en general, de elementos que poseen ciertas características comunes.

El concepto tiene acepciones específicas en diferentes ciencias naturales o sociales. En su uso más habitual, la población hace referencia a un grupo de personas que viven en un determinado lugar y así se entiende desde la demografía. Para la ecología, la población está formada por el conjunto de ejemplares de una especie que comparten un hábitat. Para la sociología, la población es un conjunto de personas, o de cosas, que pueden analizarse a partir de muestras.

Para nuestro caso particular, como investigadores educativos y sociales, la población es el conjunto de individuos o entidades que presentan las características que representan nuestro objeto de estudio. ¿Por qué ampliamos este concepto incluyendo el de "entidades"? Porque no siempre estudiamos características de individuos humanos. A veces queremos examinar, no solo estudiantes, profesores o directivos educativos, sino entidades compuestas tales como familias, escuelas o municipios, por ejemplo, o incluso entidades que ni siquiera son humanas ni están vivas, tales

como artículos publicados entre ciertas fechas, o decretos, o libros de texto, por ejemplo. Todos estos pueden ser entidades que conforman una población.

Ahora, en general hacemos investigación científica porque queremos construir conocimiento sobre una población. Por supuesto que si pudiéramos obtener información de toda la población, nuestros datos serían muy precisos. Este es el caso de estudios con poblaciones muy pequeñas. Supongamos un estudio en que queremos caracterizar la población de rectores de instituciones educativas oficiales en una entidad territorial. En este caso es más eficiente y confiable tomar datos de toda la población; esto es, hacer el estudio sobre el *censo de la población*.

Sin embargo, es muy raro que esto se pueda hacer. Muchas poblaciones son de gran tamaño y casi nunca tenemos acceso a toda la población y, aun en el caso en que lo tuviéramos, los costos implicados en recoger esta información excederían cualquier presupuesto. Lo que los investigadores hacen, casi siempre, es obtener información sobre una muestra de la población. Una *muestra de población* es una parte, más o menos pequeña, de los individuos de una población que, de alguna forma, la representan.

Así, la población es un universo desconocido que intentamos aclarar sobre la base de la información que podemos obtener de pequeñas muestras. En lo que sigue examinaremos cómo es posible que lo hagamos y cuáles son los límites que tenemos al hacerlo.

Métodos de muestreo

La capacidad que un estudio tiene para generalizar sus resultados a la población se conoce como *validez externa* del estudio. Esta resulta ser una característica fundamental de cualquier trabajo dado que habla sobre su carácter, exploratorio o no, y sobre el interés que puede tener el conocimiento construido.

En general, la validez externa de un estudio que trabaja sobre muestras de población depende de la forma de seleccionar estas muestras. Aunque existen muchos métodos para hacerlo, estos se pueden clasificar en dos grandes categorías: los probabilísticos y los no probabilísticos.

Cuando todas las unidades de la población tienen una probabilidad conocida de selección y esta es diferente de cero, se dice que la muestra es *probabilística*. En términos generales, se acepta que esta es la ideal para hacer inferencias, y muy especialmente cuando necesitamos estimar los parámetros de una población. Por ejemplo, si quisiéramos saber la capacidad general que presenta la población de grado 11 de Bogotá en relación con la resolución de problemas matemáticos, necesitaremos una muestra probabilística de esta población.

Tener una muestra verdaderamente probabilística, aunque resulta ideal, no siempre es posible. De hecho, sabemos que, incluso habiéndola diseñado de esta forma, cualquier encuesta tendrá una tasa de rechazo de cerca del 35 %, lo cual hace que, en términos estrictos, la muestra efectiva pierda el carácter de muestra probabilística y pase a ser autoseleccionada. Por esta razón, existen procedimientos para seleccionar la muestra en los casos en los que la selección no es posible hacerla con métodos probabilísticos.

Ahora, cuando en realidad no necesitamos estimar los parámetros de una población, sino solo examinar relaciones entre variables, que es un tema mucho más frecuente en la investigación

de psicología educativa, muchos investigadores desestiman la necesidad de que la muestra sea estrictamente probabilística y aceptan una muestra *por conveniencia* que tenga un cierto tamaño. Es importante recordar que, en estos casos, se debe tener precaución en la extrapolación de los resultados a la población total. Examinaremos estos dos tipos de muestreo.

Muestreo probabilístico

Los métodos de muestreo probabilísticos son aquellos que se basan en el principio de probabilidad, lo que significa que todos los individuos tienen una probabilidad conocida de ser elegidos para formar parte de una muestra. Los métodos de muestreo probabilístico son los únicos que nos aseguran la representatividad de la muestra extraída y son, por tanto, los más recomendables.

Dentro de los métodos de muestreo probabilístico encontramos los siguientes tipos:

- *Muestreo aleatorio simple:* todos los elementos tienen la misma probabilidad de ser seleccionados, y la selección se hace por números aleatorios. Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población que estamos manejando es muy grande.
- *Muestreo aleatorio sistemático:* se trata de seleccionar muestras cuando la población ya está dividida. Por ejemplo, en un colegio, se pueden tomar uno de cada cinco individuos en cada curso, según el orden en el que aparezcan en la lista.
- *Muestreo aleatorio estratificado:* consiste en considerar categorías típicas diferentes entre sí (llamados estratos), que poseen gran homogeneidad respecto a alguna característica que resulta de interés para el estudio. Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra. Cada estrato funciona de manera independiente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple. En ocasiones las dificultades que plantea este tipo de muestreo son demasiado grandes, pues exige un conocimiento detallado de la población. Un ejemplo puede ser tomar las instituciones educativas según su categoría de rendimiento en las pruebas de estado (muy superior, superior, etc.).
- *Muestreo aleatorio por conglomerados:* los métodos presentados hasta ahora están pensados para seleccionar directamente los elementos de la población, es decir, las unidades muestrales son los elementos de la población. En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que llamamos "conglomerado". Los grupos escolares, por ejemplo, son conglomerados naturales de estudiantes en la investigación educativa. Este muestreo consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y en investigar después todos los elementos pertenecientes a estos. El análisis puede hacerse en varias etapas: por ejemplo, primero se seleccionan colegios, después sedes y por último cursos (esto se conoce como *muestreo polietápico*). Este tipo de muestreo es más sencillo a la hora de administrar el proceso de recolección de la información, ya que todo el conglomerado informa y no tenemos que separar algunos individuos. Por esta misma razón se puede recoger mayor cantidad de información en menos tiempo.

Muestreo no probabilístico

A veces, para estudios exploratorios o para estudios de pruebas de hipótesis, el muestreo probabilístico resulta excesivamente difícil o costoso y se acude a métodos no probabilísticos. En algunas circunstancias, los métodos estadísticos y epidemiológicos permiten resolver los problemas de representatividad aun en situaciones de muestreo no probabilístico, como por ejemplo los estudios de caso-control, en los que los casos no son seleccionados aleatoriamente de la población.

Entre los métodos de muestreo no probabilísticos más utilizados en investigación encontramos:

- *Muestreo intencional o de conveniencia.* Este tipo de muestreo se caracteriza por un esfuerzo deliberado de obtener muestras "representativas", mediante la inclusión en la muestra de grupos supuestamente típicos. En este tipo de muestreo es usual que el investigador seleccione directa e intencionadamente los individuos de la población. El caso más frecuente de este procedimiento es el de utilizar, como muestra, los individuos a los que se tiene fácil acceso (por ejemplo, los profesores de universidad emplean a menudo a sus propios alumnos).
- *Muestreo por cuotas.* También denominado en ocasiones "accidental". Se asienta generalmente sobre la base de un buen conocimiento de los estratos de la población o de los individuos más "representativos" o "adecuados" para los fines de la investigación. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquel. En este tipo de muestreo se fijan unas "cuotas" que consisten en un número de individuos que reúnen unas determinadas condiciones, por ejemplo, cinco escuelas en un departamento. Una vez definida la cuota, se eligen los primeros que se encuentren que cumplan esas características.
- *Bola de nieve*. Se localizan algunos individuos, los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente. Este tipo de muestreo se emplea muy a menudo cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas, determinados tipos de enfermos, etc., a los que no se tiene fácil acceso.
- *Muestreo discrecional.* A criterio del investigador los elementos son elegidos sobre lo que él cree que pueden aportar al estudio.

El problema del tamaño de muestra

La historia de la estadística narra que una de las grandes diferencias entre dos de los fundadores de la estadística, K. Pearson y R. Fisher, se basó en sus opiniones sobre tamaño de muestra para la toma de decisiones. Pearson era partidario de tamaños de muestras grandes; cuando Gosset (quien publicó bajo el seudónimo de Student) le presentó los datos de su trabajo sobre la media muestral en muestras pequeñas, él lo apoyó para su publicación en *Biométrica*, pero no le dio mayor importancia. Fue Fisher quien le dio la importancia, apoyó a Gosset y corrigió la distribución *t*.

Cuando intentamos hacer una inferencia estadística, y muy especialmente cuando estamos estimando un parámetro de población, el tamaño de la muestra es muy importante, ya que sabemos que la precisión de una estimación va aumentando a medida que aumenta el tamaño de la muestra.

Se llama *error de muestreo* a la diferencia entre el valor muestral del estimador y el parámetro que se estima. Por supuesto, cuanto menor sea el error de muestreo, mayor será la precisión del estimador. El punto es que, cuanto mayor sea el tamaño de la muestra, menor será el error de muestreo.

Como los parámetros por estimar (medias, proporciones, tasas o totales) son estadísticos (es decir, son calculados sobre muestras), entonces, tienen su propia distribución de probabilidad. Por esta razón se define, en general, el *error estándar* como la desviación estándar de la distribución muestral del estadístico. Cuanto más pequeño sea el error estándar, más preciso será el estadístico y, por lo tanto, más cercano será al parámetro poblacional.

Ya antes, cuando estudiamos las medidas de dispersión, habíamos definido el error estándar de la media (EEM) como la relación entre la desviación estándar muestral y la raíz cuadrada del tamaño de la muestra. Como se observa, este es un caso particular del concepto general de error estándar.

Cada diseño muestral tiene su propio método para calcular el estimador y su propio método para calcular el error estándar. La medida usual para interpretar esta precisión es el coeficiente de variación (CV), que ya estudiamos antes. Una guía internacional para la interpretación de este indicador es la que se observa en la tabla 36.

Tabla 36. Interpretación del coeficiente de variación

Menos de 10 %	Estimación precisa
Entre el 10 % y el 14 %	Precisión aceptable.
Entre el 15 % y 20 %	Precisión regular. Se debe utilizar con precaución.
Mayor del 20 %	La estimación es poco precisa. Se recomienda utilizarla solo con fines descriptivos.

Por otra parte, en los estudios de pruebas de hipótesis, el problema no es el de la estimación de parámetros poblacionales, sino el del *cálculo de la probabilidad* de que un suceso o un evento (una asociación o una diferencia) se encuentre presente en la población muestreada. Para ello no se requiere de un gran tamaño de muestra; es más importante seleccionar la distribución apropiada para calcular la probabilidad de ocurrencia del evento observado.

Las pruebas no paramétricas, por ejemplo, están diseñadas para muestras menores a treinta individuos. En las correlaciones de Spearman o Kendall, por ejemplo, pueden considerarse muestras de tamaños tan pequeños como seis individuos; así lo atestiguan las tablas que describen estas probabilidades e inician con ese valor (Siegel, 1980).

En las pruebas paramétricas se requieren tamaños de muestra un poco mayores. En la prueba *t* de Student para grupos independientes, por ejemplo, utilizada para comparar medias de dos grupos, deben contarse con, al menos, seis sujetos en cada grupo; esto es, no menos de doce sujetos en total. En el análisis de varianza de un factor, las tablas inician con un tamaño de muestra mínimo de quince individuos.

En general, debe anotarse que el tamaño de la muestra afecta aspectos relacionados con el tamaño del efecto y la potencia de la prueba estadística. Por esta razón, una diferencia pequeña que no parece ser significativa en una muestra pequeña, sí pudiera serlo en una muestra de gran tamaño.

Esto representa la otra cara de la moneda: diferencias menores pueden parecer muy importantes si la muestra es de tamaño muy grande. Más adelante se presentará y se desarrollará el tema de la potencia estadística y el tamaño del efecto, por lo que deberemos posponer esta discusión.

Formas en que se expresa la información sobre muestras en publicaciones científicas

La media, la desviación estándar y la varianza de una población, en su totalidad, se denominan *parámetros poblacionales* y se simbolizan mediante las letras griegas: μ (media), σ (desviación estándar) y σ^2 (varianza). Estos parámetros rara vez se conocen y usualmente se estiman a partir de muestras de población cuidadosamente seleccionadas.

Por su parte, la media, la desviación estándar y la varianza que calculamos con base en los datos de una muestra se conocen como *estadísticos muestrales* y se simbolizan con las letras *M* o para la media, *DE* para la desviación estándar (*SD* en inglés) y DE^2 para la varianza (SD^2, en inglés).

En general, el tamaño total de la muestra (número de casos) se denota con la letra *N*. Esta misma letra en minúscula (*n*), o subindicada (n_1, n_2…) se utiliza para denotar el número de casos en submuestras.

En situaciones de proyectos en los que se desea estimar parámetros poblacionales, el reporte de la calidad de la muestra resulta imprescindible. Para hacerlo, se reporta el estimador, su error estándar y su coeficiente de variación. En estudios de pruebas de hipótesis, por el contrario, basta con mencionar los tamaños de muestra, totales y por subgrupos, y los tamaños observados del efecto. En algunos de ellos se reporta también la potencia de la prueba, pero esto no resulta tan frecuente.

Probabilidad

Concepto

En general, cuando adelantamos una investigación educativa o social tenemos como objetivo examinar las predicciones que se hacen a partir de una teoría (lo que llamamos someter a *falsación* esta teoría), o bien examinar los efectos de algún tipo de intervención sobre la población. Por supuesto, no podemos trabajar sobre toda la población y, por tanto, es imposible que podamos probar, con el 100 % de certeza, que esta teoría es verdadera, o que la intervención examinada siempre tiene tal o cual efecto. En vez de ello, lo que nos permite la estadística inferencial es conocer, con cierta precisión, la *probabilidad* de que el resultado que encontramos siga siendo el mismo si lo obtuviéramos en toda la población. Por esta razón, el concepto de probabilidad es crucial para nosotros.

En estadística, se define la *probabilidad* como la frecuencia relativa con la cual esperamos que suceda determinado resultado; dicho de otra forma, la proporción de resultados esperados dentro del total de resultados posibles. Por ejemplo, ¿cuál es la probabilidad de que un dado caiga en el número uno al ser lanzado? El dado puede dar seis resultados, pero nuestro resultado esperado es uno, por lo que la probabilidad será 1/6, o 0,1666... o, expresado en términos porcentuales, aproximadamente el 16,66 %.

En general, expresamos las probabilidades como números entre 0 y 1. Cuando un hecho no puede ocurrir, o es *imposible*, diremos que tiene una probabilidad de cero. En el otro extremo, cuando un evento tiene una probabilidad de uno, diremos que es un evento *seguro*. Si un evento tuviera una probabilidad baja, digamos del 10 % (0,1), o del 5 % (0,05) o, más baja aún, del 1 % (0,01) diremos que es un evento *improbable*. Es importante recordar que independientemente de qué tan improbable sea un evento, si este no es imposible, podría ocurrir.

Esta definición de probabilidad, dada en términos de frecuencia relativa, es conocida como *definición frecuentista de probabilidad*. Como una alternativa a esta definición, en los últimos años ha surgido toda una corriente en la prueba de hipótesis conocida como *probabilidad bayesiana*. Esta corriente, fundamentada en el teorema de Bayes, se caracteriza porque intenta tener en cuenta los resultados previos para actualizar las estimaciones de los resultados de la nueva muestra.

El objetivo de la estadística bayesiana es cuantificar la incertidumbre anexa a la inferencia. Así, la probabilidad bayesiana trata los parámetros como variables aleatorias que pueden describirse con una distribución de probabilidad. Las estimaciones se ponderan mediante el factor de actualización predictiva, y combina la información previa con los nuevos resultados. Esta relación de rendimiento predictivo se conoce como *factor de Bayes*. Como se entiende, esta metodología puede estar afectada por elementos subjetivos del investigador o por el hecho de que los resultados previos sean producto de una metodología no compatible con el nuevo diseño.

Aunque las pruebas de hipótesis han usado, tradicionalmente, una aproximación frecuentista, cada vez es más fácil ver alternativas bayesianas en los programas de procesamiento estadístico. En los dos programas que utilizamos en este trabajo se plantean estas dos alternativas. El JASP ofrece, desde sus inicios, para prácticamente todos los procedimientos, alternativas bayesianas. En el caso del IBM-SPSS no se brindaba esta alternativa, pero ya en las últimas versiones (a partir de la v. 26) presenta procedimientos de orientación bayesiana.

Formas en que se expresa la probabilidad en publicaciones científicas

En general, la probabilidad se simboliza con la letra *p*, en cursiva. Existen dos convenciones importantes, en el formato APA, para expresar la probabilidad de un suceso.

- La primera consiste en que, cuando un número no puede superar el valor 1, como es el caso de la probabilidad, al expresarlo se omite el dígito "0" de los enteros. Utilizando esta convención, si queremos expresar una probabilidad del 50 %, no la expresaremos como p=0,5, sino como p=,5 (en inglés no utilizamos la coma decimal, sino el punto decimal: p=.5).
- La segunda se refiere al número de decimales. En general, se sugiere que todas las medias se expresen con dos decimales, a excepción de la probabilidad, que debe ser expresada con tres decimales.

Rangos de probabilidad de uso muy frecuente en la investigación social, para expresar sucesos muy poco probables son p<,05 (probabilidad menor al 5 %) o p<,01 (probabilidad menor al 1 %) o, mejor aún, p<,001 (probabilidad menor al uno por mil, o 0,1 %).

Distribución normal

Los histogramas de muchas de las variables que estudiamos en la investigación educativa y, en general, en la investigación social tienen una forma característica, que tiende a ser simétrica y unimodal, como una campana. Este tipo de distribución es bien conocida en la estadística y se le denomina *distribución normal* o *curva normal*. Otros nombres para esta distribución son *distribución de Gauss* o *campana gaussiana* en honor al célebre matemático alemán Johann Carl Friedrich Gauss (1777-1855) quien, dicho sea de paso, no fue el descubridor de la curva normal. Su descubrimiento se atribuye al matemático francés Abraham de Moivre (1667-1754).

La gráfica de la figura 38 muestra una curva normal estándar. La fórmula para su cálculo, cuando consideramos una media de 0 y una desviación estándar de 1, es

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

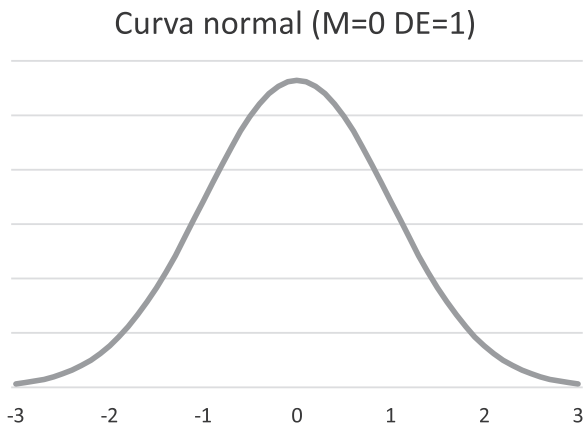


Figura 38. Curva normal (M=0 y DE=1)

Las observaciones aproximadamente normales son frecuentes en la naturaleza. Si aplicamos un instrumento de evaluación a una población suficientemente grande y relativamente homogénea, y contamos el número de ítems correctamente resueltos, notaremos que la mayoría de los resultados se agrupan de forma simétrica alrededor de la media. Esto crea una distribución simétrica y unimodal. Sin embargo, puede demostrarse matemáticamente que, si las circunstancias fueran al azar, el resultado sería una perfecta curva normal. Este se conoce como el *teorema del límite central*.

La distribución normal es bastante usada en problemas en los que la simetría es relevante. Variables como el peso en una población de niños de cierta edad, sus estaturas, rendimientos académicos o puntajes en pruebas, entre otras, se comportan de forma normal.

Dado que la curva normal es estándar y la conocemos bien, es posible saber el porcentaje de casos que se encuentra entre dos valores determinados. Dado que es una curva perfectamente simétrica, sabemos que el 50 % de los casos se ubica por encima de la media, y el 50 % por debajo. Otros puntos de referencia importantes son la media más o menos una o dos desviaciones estándar. La gráfica de la figura 39 muestra estos porcentajes.

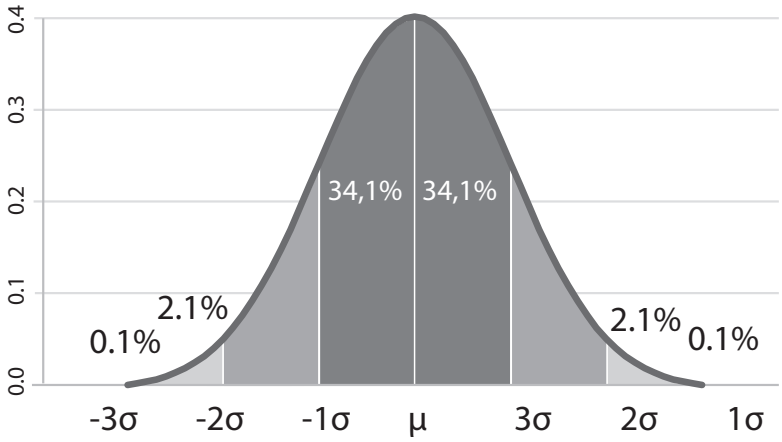


Figura 39. Curva normal con porcentaje de casos bajo la curva

Así, sabemos que el 68,2 % de los casos se ubican alrededor de la media más, o menos, una desviación estándar, y que aproximadamente el 95,8 % de los casos se ubican alrededor de la media y a una distancia no mayor de dos desviaciones estándar. Esto nos señala también que es bastante improbable, aunque no imposible, que, al escoger un caso al azar, este se ubique a más de dos desviaciones estándar de la media.

Esto nos conecta dos conceptos, hasta ahora aislados: el de probabilidad y el de la curva normal. La distribución normal puede ser considerada una *distribución de probabilidades*. La proporción de valores entre dos puntuaciones Z cualquiera es exactamente lo mismo que la probabilidad de seleccionar un valor cualquiera entre estos dos valores Z. En concreto, sabemos que la probabilidad que tenemos de que, al seleccionar un valor al azar, este se encuentre a menos de una desviación estándar de la media es del 68,2 %: relativamente alta, diríamos.

Existen otras distribuciones de probabilidad utilizadas para el cálculo de probabilidades en eventos específicos, tales como las distribuciones muestrales continuas t, F o χ^2 (chi-cuadrado) y las distribuciones discretas de probabilidad binomial y multinomial. Esas distribuciones serán utilizadas más adelante en las pruebas de hipótesis que las toman como base.

Capítulo 8

La prueba de hipótesis

Teoría e hipótesis

Antes de poder ser considerada como una explicación científica correcta, una teoría debe 1) tener sentido y mejorar la comprensión de los fenómenos, y 2) aportar predicciones empíricamente verificables (esto es, predicciones observables y medibles). Relacionar el conjunto de ideas y conceptos de una teoría con lo que se debería observar en el mundo real, si tal teoría se verificara, es un momento fundamental del desarrollo científico. Para este momento, la estadística hace un importante aporte proporcionando un método objetivo para la prueba de hipótesis.

Primero, es necesario imaginarse el futuro en términos de la formulación de una hipótesis. Una *hipótesis* es una predicción que requiere comprobación por medio de la observación y el análisis de datos. La palabra hipótesis, que comparte su raíz con el término *hipotético*, es una conjetura científica que requiere contrastación por la experiencia; es un enunciado no verificado aún; en cuanto es confirmado, o refutado, deja de ser una hipótesis.

Ahora, cuando formulamos una hipótesis, nos referimos a un hecho que, suponemos, se da para *todos* los miembros de una población determinada. Sin embargo, a la hora de comprobarla, con frecuencia no tenemos acceso a datos de toda la población y debemos conformarnos con los datos que podemos obtener de una pequeña muestra. Las formas de generalización de lo observado en una muestra para la población total son el campo de la estadística inferencial.

Dicho de otra forma, una hipótesis estadística es una afirmación, o conjetura, que se hace respecto a una o más poblaciones desde uno o varios parámetros poblacionales. La prueba de hipótesis es un procedimiento sistemático para determinar si los resultados de una observación hecha sobre una muestra de población son válidos en relación con lo que se observaría en la totalidad de la población. La validez de esta hipótesis se establecerá de forma probabilística, es decir, nunca se llega a saber con exactitud, ya que se parte de datos de una muestra.

Para poder explicar todos los conceptos asociados a la prueba de hipótesis, plantearemos y desarrollaremos un ejemplo concreto, que se presentará en recuadros en lo que sigue del capítulo.

Recuadro 20. Planteamiento de un ejemplo. Relaciones entre dos pruebas de logro

Ilustraremos la prueba de hipótesis con un ejemplo sobre las relaciones entre los resultados de dos pruebas: Lenguaje y Ciencias Naturales en una pequeña muestra de estudiantes de grado 10.

Un investigador tiene la opinión de que los resultados obtenidos por los estudiantes en una prueba de Ciencias tienen una fuerte relación con las competencias de los estudiantes en la lectura e interpretación de los textos de la prueba. Para verificar esta suposición aplica dos pruebas: una de Ciencias y otra de Lenguaje en una muestra de estudiantes.

En este caso, contamos con una hipótesis clara: cuando calculemos la correlación entre los resultados de la prueba de Lenguaje y los de la prueba de Ciencias en nuestra muestra, encontraremos una alta correlación.

Supongamos que, efectivamente, aplicamos las dos pruebas y calculamos la correlación entre estas dos variables. Los resultados podrían ser de dos tipos. Primero, puede ser que la correlación sea muy baja, muy cercana a cero. En este caso, los resultados son claros, y podemos rápidamente concluir que, con la metodología que se utilizó, no parece haber una relación lineal clara entre estas dos variables.

Segundo, es posible que encontremos que la correlación entre las dos pruebas muestre ser relativamente alta, ya sea en el sentido positivo o en el sentido negativo. ¿Esto permitiría concluir que se comprueba la hipótesis de que las dos variables están muy relacionadas? No necesariamente. Es posible, por ejemplo, que encontráramos que, aunque la correlación no es cero, no es demasiado distante: podría ser 0,2, por ejemplo. Una correlación de este nivel podríamos encontrarla, con cierta frecuencia, por pura casualidad, entre variables que no están relacionadas. ¿Qué nos asegura que no tenemos tan buena, o tan mala, suerte que precisamente encontramos una muestra donde estas dos variables estén relacionadas por puro azar?

En otras palabras, lo que nos preguntamos ahora es ¿en qué medida los resultados de nuestra observación pueden deberse al hecho de que trabajamos con una muestra y no con la totalidad de la población? O, mejor, ¿en qué medida los resultados pueden ser producto del error muestral? En términos prácticos, ¿qué tan alta debe ser la relación, para que aceptemos la hipótesis? Para responder esto, debemos adentrarnos en la lógica de las pruebas de hipótesis.

La lógica de las pruebas de hipótesis

Para responder a la pregunta original, debemos calcular la correlación lineal entre estos dos resultados, tal y como la examinamos en el capítulo 4. Si encontráramos que la correlación entre estas dos variables fuera muy pequeña, muy cercana a cero, por ejemplo, tendríamos que concluir que, con alta probabilidad, estas dos variables son independientes entre sí, y que el resultado de una no contribuye a predecir el resultado de la otra. Esto no sería para nada extraño: muchas variables no tienen relaciones entre sí.

Ahora, si este no fuera el caso, sino que, por el contrario, encontráramos que la correlación entre estas dos variables en nuestra muestra fuera muy alta, ¿cómo podríamos explicarlo? Primero, tendríamos que considerar que encontrar una correlación muy alta es muy difícil por puro efecto del azar. Si sabemos que es muy improbable que la encontráramos así por puro azar, entonces,

tendríamos bases para rechazar la idea de que esta alta correlación apareció allí por azar. Si debemos rechazar la idea de que esto fue por azar, debemos entonces, aceptar que, seguramente, este no fue un resultado fortuito, sino que esta alta correlación en nuestra pequeña muestra representa lo que efectivamente ocurre en la población total.

En otras palabras, si encontráramos que la correlación entre las dos variables fuera tan alta que resultara muy improbable encontrarla por azar, tal vez deberíamos considerar la posibilidad de que esta correlación sea explicable por alguna forma de relación efectivamente presente entre estas dos variables: tal vez una sea la causa de la otra, o tal vez en las dos se requiere una habilidad común, por ejemplo; no podemos saberlo. Solo sabríamos que nuestro resultado está muy probablemente presente en la totalidad de la población. Esa es la lógica de la prueba de hipótesis.

Afortunadamente, los estadísticos han calculado la probabilidad de obtener determinado coeficiente de correlación (r) en una muestra de cualquier tamaño por efecto del azar, para lo cual han utilizado las distribuciones t y normal de probabilidad. En términos generales, la probabilidad p, de obtener un coeficiente determinado de correlación r en una muestra de tamaño n es más baja cuanto más altos sean r y n. En otras palabras, altos coeficientes de correlación son muy improbables en muestras grandes.

Supongamos, para continuar con nuestro ejemplo, que obtuvimos un coeficiente de correlación muy alto que podría ocurrir, por efecto del azar en menos del, digamos, 2 % de los casos. Esta probabilidad es muy baja, sin duda.

Como esta probabilidad es tan baja, podemos asumir que esto no fue lo que ocurrió, es decir, este evento muy probablemente no ocurrió por efecto del azar. Si no ocurrió por azar, muy posiblemente sí ocurrió por el hecho de que esta asociación está presente en la totalidad de la población. En este caso, habríamos probado la hipótesis de que las dos variables están asociadas significativamente.

Dicho en otra forma, una probabilidad tan pequeña de obtener este resultado por azar nos sugiere *rechazar* la idea de que las dos variables *no* están relacionadas. Si debemos rechazar la idea de que las dos variables no están relacionadas, debemos aceptar la idea de que sí lo están. Esa es la lógica de la prueba de hipótesis.

Este tipo de razonamiento, al revés, es la clave de la inferencia en estadística. Es una doble negación en la que hemos determinado la probabilidad de obtener el evento contrario al que estamos prediciendo. Si esta probabilidad es muy baja, entonces, concluimos apoyando la hipótesis de que es inverosímil que tal evento no ocurra.

El proceso de la prueba de hipótesis

El proceso de la prueba estadística de hipótesis es producto de discusiones y aportes de los principales estadísticos del siglo xx. Para decidir, de manera objetiva, si una hipótesis teórica es confirmada por un conjunto de datos, es necesario un procedimiento con criterio objetivo para aceptar, o rechazar, esta hipótesis a partir de una muestra seleccionada con este propósito.

Es posible diferenciar varios momentos fundamentales en el proceso de prueba de hipótesis. De acuerdo con un punto de vista actual, en el que se asume que se cuenta con un buen *software* para el procesamiento y análisis de datos, el proceso podría ser esquematizado como sigue:

1. Formulación de la hipótesis. En este paso, se hacen explícitas las hipótesis que esperamos contrastar con nuestros datos. Deberemos decir que existe una relación, o que existe una diferencia entre un cierto número de medidas y, tal vez, deberemos decir cómo es la relación o la diferencia.
2. Selección de la prueba adecuada. La prueba dependerá de la hipótesis que se desea corroborar, del tipo de medida de las variables involucradas y del cumplimiento de ciertas condiciones o "supuestos" que son propios de cada prueba. Algunas pruebas requieren el cumplimiento de ciertos supuestos, otras no.
3. Cálculo de los estadísticos, los niveles de significación y el tamaño del efecto. Seleccionada la prueba y verificados sus supuestos propios, puede ejecutarse la prueba en el *software*.
4. Interpretación y expresión de los resultados. Los resultados arrojados por el programa deben ser interpretados y expresados en una forma completa y comprensible.

Explicaremos en detalle este proceso utilizando para ello el ejemplo del investigador que sospecha relaciones muy estrechas entre una prueba de Ciencias y una prueba de Lenguaje.

Paso 1. Formulación de la hipótesis

Hipótesis nula y alternativa

Para poder aplicar la prueba, se formulan dos tipos de hipótesis:

> *Hipótesis nula (H_0). Hipótesis de nulidad o hipótesis de diferencias nulas. Este tipo de hipótesis expresa la ausencia de relación, diferencia, causalidad, etc. entre dos o más variables. Se suele esperar rechazarla después de la prueba estadística.*
>
> *Hipótesis alternativa (H_1). Hipótesis de la investigación. Plantea una aseveración, conjetura o proposición sobre las probables relaciones entre dos o más variables. Con frecuencia se puede expresar en forma descriptiva, correlacional o de causalidad, dependiendo del propósito y de la naturaleza de la investigación. Es la afirmación que se espera aceptar después de aplicar la prueba.*

Hipótesis unilaterales o bilaterales

En términos generales, existen dos formas de plantear la hipótesis dependiendo de si se supone una cierta y concreta dirección en el resultado esperado o no, se hace esta suposición. Específicamente, se puede decir que, entre dos variables dadas existe una relación y que la relación entre estas dos variables es estrictamente positiva (esto es, no negativa) o, por el contrario, se puede decir que estas dos variables están relacionadas, de alguna forma (positiva o negativa).

Las hipótesis en las que suponemos una específica dirección del efecto se conocen como *hipótesis direccionales*, *unidireccionales* o *unilaterales*. Cuando se examina este tipo de hipótesis, la probabilidad de encontrar un determinado resultado debe mostrar un valor que se ubique dentro del 5 % superior de la distribución. Si el valor se encontrara, por el contrario, dentro del 5 % inferior, no podría descartarse la hipótesis nula. Así, para una hipótesis unidireccional se utilizaría una *prueba de una cola.*

En muchos casos, la hipótesis de investigación se formula de manera que indica que existe una relación entre las dos variables, sin especificar si esta relación será positiva o negativa. Este segundo tipo de hipótesis se conoce como hipótesis *no direccional, bidireccional* o *bilateral*. En estos casos, el investigador debe examinar si los valores del grupo experimental se encuentran en cualquiera de las dos colas: la superior o la inferior. Por esta razón, en estos casos se aplica una *prueba de dos colas*.

¿Qué implicaciones tiene esto? Muchas, en realidad. La más importante es que una hipótesis unidireccional es menos exigente que una bidireccional, en el sentido en que es más fácil rechazar la hipótesis nula (y por tanto, aceptar la hipótesis alternativa) con una prueba de una cola. En efecto, con hipótesis de una cola los niveles de correlación encontrados no tienen que ser tan extremadamente altos para poder rechazar la hipótesis nula. El problema es que, si llegásemos a encontrar, de forma completamente inesperada y contradictoria con lo que buscamos, relaciones que contradigan flagrante y decididamente la suposición inicial, tampoco es posible desechar la hipótesis nula.

Esto ocurre porque el 5 % de los resultados que consideraremos significativos en una prueba de una cola se ubican en una sola dirección de la curva. En la prueba de dos colas, este 5 % debe distribuirse, la mitad en la cola inferior y la mitad en la cola superior, haciendo que los valores límites sean más extremos que los observados en la prueba de una cola.

En la práctica, la mayoría de los investigadores utilizan pruebas de dos colas, tanto para hipótesis unidireccionales como para las bidireccionales. Si el resultado de la prueba de dos colas es significativo, entonces, se considera el resultado significativo en la dirección encontrada. Este procedimiento, si se piensa, es bastante riguroso, ya que es más difícil encontrar un resultado significativo en una prueba de dos colas por lo que, en el caso en el que se obtenga, se da más seguridad al resultado. En términos generales y, a menos que se especifique lo contrario, se supone que se utilizan pruebas de dos colas.

Ejemplo. Relaciones entre dos pruebas de logro

Recuadro 21. Ejemplo Relaciones entre dos pruebas de logro (continuación).

1. Formulación de las hipótesis

Para ilustrar el proceso, desarrollaremos un ejemplo. Supongamos que un investigador tiene la idea de que los resultados obtenidos por los estudiantes de grado 10 en una institución educativa en una prueba escrita que examina los conocimientos en ciencias naturales podrían ser, en buena parte, explicables por las capacidades que este grupo muestra para leer de forma comprensiva. Si los estudiantes no consiguen comprender el texto que acompaña la prueba de Ciencias, fallarán en la prueba, razona el investigador, no por sus conocimientos de ciencias, sino por su comprensión de lectura.

Para probar esta idea el investigador deberá hacer muchos arreglos. Como una primera aproximación al tema, puede aplicar pruebas de Lenguaje y de Ciencias Naturales a un mismo grupo escolar de grado 10 compuesto por treinta estudiantes, con la idea de que los resultados de estas dos pruebas podrían estar muy relacionados.

Para el caso de nuestro ejemplo, referente a las relaciones entre los resultados de las pruebas de Lenguaje y de Ciencias, podemos plantear las hipótesis de la siguiente forma:

- Hipótesis nula (H_0). No hay una relación lineal entre los resultados de la prueba de Lenguaje y los de la prueba de Ciencias Naturales.
- Hipótesis alternativa (H_1). Existe una relación lineal entre los resultados de la prueba de Lenguaje y los de la prueba de Ciencias Naturales

Como se observa, las hipótesis son definidas como bidireccionales. Mantendremos esta convención en lo que sigue.

Paso 2. Selección de la prueba estadística adecuada

En términos generales, la selección de la prueba tiene dos momentos: el primero es el de una preselección de la prueba; en el segundo, verificamos sus supuestos propios y tomamos decisiones al respecto.

Preselección de la prueba

Actualmente, los investigadores cuentan con una gran cantidad de pruebas estadísticas apropiadas para múltiples propósitos y situaciones. La selección de la prueba depende del tipo de hipótesis que se pretende contrastar, del nivel de medida de las variables, de la naturaleza de la población y del tamaño de la muestra. Todos estos factores deben examinarse y la prueba seleccionada deberá cumplir con las condiciones o supuestos que le son propios.

En el capítulo 9, haremos una presentación general de las diferentes pruebas estadísticas disponibles para las situaciones más comunes. Por otra parte, en los últimos capítulos de esta obra examinaremos en detalle algunas de las pruebas de uso más común en la investigación educativa y social.

Ejemplo (continuación). Preselección de la prueba

Recuadro 22. Relaciones entre dos pruebas de logro (continuación)

2a. Elección de la prueba estadística adecuada.

Para el caso de nuestro ejemplo, tenemos una situación en la que se desea examinar la relación lineal entre dos variables cuyo nivel de medida es métrico. La prueba ideal para este caso, como lo expusimos en el capítulo 4, es *la correlación producto-momento de Pearson*.

Verificación de supuestos

Una vez preseleccionada la prueba, se pasa a verificar cuáles son los supuestos necesarios para aplicarla. Todas las pruebas estadísticas presentan una serie de *supuestos*, o condiciones necesarias para su aplicación. En algunas pruebas, la violación de algunos de los supuestos afecta gravemente el alcance de la inferencia; en otras puede no ser tan grave o, incluso, admitirse la violación

moderada de algún supuesto. Esto depende de la prueba en cuestión. Más adelante, al finalizar el capítulo 9, examinaremos los supuestos más comunes para la mayoría de las pruebas y las formas adecuadas de verificarlos o de solucionar el problema de su no verificación.

Existe la posibilidad de que la prueba preseleccionada no cumpla con uno o varios de sus supuestos propios. En este caso, debe procederse en una secuencia ordenada de cuatro posibles soluciones: 1) la valoración del incumplimiento del supuesto en la prueba específica; 2) la exploración de la existencia de "correcciones" a la prueba para el caso del incumplimiento del supuesto; 3) la ejecución de transformaciones a las variables para hacer que se cumpla el supuesto y, si todo lo anterior falla, 4) la selección de una nueva prueba, usualmente una alternativa no paramétrica equivalente a la original, que no requiera del cumplimiento del supuesto.

La primera solución es sencilla. Debe ser examinado el supuesto incumplido y su efecto en la prueba específica. Para algunas pruebas, la violación moderada de algún supuesto específico no es demasiado grave; se dice en estos casos que esa prueba es "robusta" frente a la violación de ese supuesto. En el análisis de varianza en una dirección (Anova *one way*), por ejemplo, se requiere el cumplimiento del supuesto de normalidad de la variable dependiente, pero se admiten casos de variables relativamente simétricas sin valores atípicos. Si este es el caso, podemos proceder directamente con la prueba en cuestión. Para otras pruebas, la violación de algún supuesto es un punto crítico y podría afectar seriamente la confianza en los resultados. Aquí deberá procederse a examinar la segunda solución.

La segunda solución depende de los desarrollos que se hayan hecho de la prueba específica. Para algunas pruebas, se han desarrollado formas específicas (llamadas "correcciones") que se aplican en el caso de violaciones a supuestos importantes. Este es el caso, por ejemplo, de la corrección de Welch, utilizada cuando se viola el supuesto de igualdad de varianzas en una prueba *t* de Student. Este tipo de correcciones soluciona el problema de la violación de algún supuesto modificando levemente la prueba y sus niveles de significación. Existen diferentes correcciones para el caso de la violación de algunos supuestos en diferentes pruebas. Cuando estas alternativas existen, son las soluciones más fáciles a la violación de un supuesto específico y deben ser utilizadas. Cuando estas correcciones no existen, deberemos proceder a la tercera solución.

La tercera solución consiste en transformar las variables que no están cumpliendo con los supuestos exigidos por la prueba. Este procedimiento consiste en aplicar alguna función a la variable que no cumple el supuesto de manera que la variable transformada lo cumpla. Las funciones más ampliamente utilizadas para este efecto son la raíz cuadrada o el logaritmo. Si se utilizan transformaciones y la variable transformada permite el cumplimiento del supuesto, la prueba deberá ser corrida sobre la variable transformada y ello deberá ser considerado a la hora de interpretar los resultados. Al final del capítulo 9 se examinarán las transformaciones más comunes para la resolución de los incumplimientos típicos.

Existen casos de violaciones importantes en los supuestos críticos en los cuales no hay correcciones y no es posible encontrar transformaciones que solucionen esta situación. En estos casos, el investigador deberá proceder a la cuarta de las soluciones: seleccionar la prueba estadística no paramétrica equivalente a la prueba preseleccionada. Para muchas de las pruebas paramétricas existe un equivalente no paramétrico que, por su naturaleza, es libre de distribuciones y de los supuestos de normalidad y homocedasticidad que las pruebas paramétricas requieren. Por ejemplo, el

equivalente a la prueba *t* de Student para grupos independientes es la prueba *U* de Mann Whitney; el equivalente no paramétrico al análisis de varianza de una vía es la prueba *W* de Kruskal-Wallis. En estos casos extremos, la alternativa no paramétrica constituye una buena solución al problema.

El diagrama de flujo de la figura 40 representa el árbol de decisiones que hemos descrito.

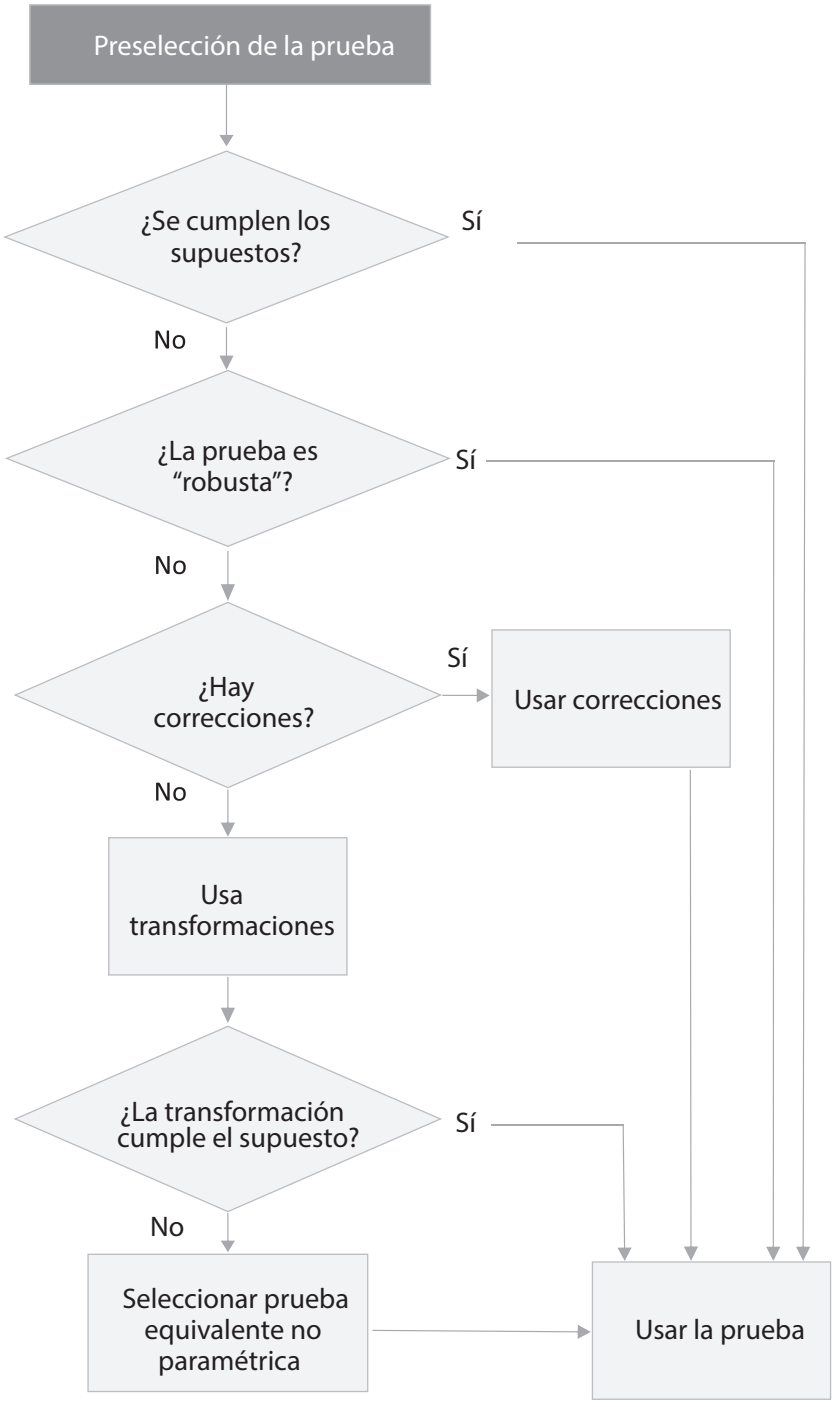


Figura 40. Diagrama de flujo para la verificación de supuestos de una prueba

Ejemplo (continuación): verificación de supuestos

Recuadro 21. Ejemplo. Relaciones entre dos pruebas de logro (continuación)

2b. Verificación de los supuestos de la prueba

En el caso de nuestro ejemplo, debemos recordar que estamos examinando una relación lineal entre dos variables métricas. Considerando estas características, hemos definido una prueba de asociación, tal como el coeficiente de correlación producto-momento de Pearson. El sentido de este coeficiente y sus formas de interpretación ya fueron expuestos en el capítulo 3. Lo que faltaría en este punto es asociarlo con un determinado nivel de significación.

Específicamente, la prueba del coeficiente de correlación producto-momento de Pearson tiene los siguientes supuestos o condiciones:

Las unidades deben haber sido seleccionadas aleatoriamente.

1. Nivel de medida: las dos variables deben ser métricas.
2. Linealidad. La relación entre las dos variables debe ser lineal o, al menos, no debe ser no lineal.
3. Normalidad. Las dos variables deben distribuirse normalmente.
4. Normalidad bivariada. Este presupuesto supone que las dos variables se distribuyen de forma normal, de manera conjunta; en otras palabras, cada variable debe distribuirse normalmente en cada valor de la otra variable.

Debemos verificar los supuestos para la aplicación de la prueba de correlación de Pearson. Los primeros dos supuestos están dados por la forma de selección de la muestra y por la naturaleza de las medidas; en este caso, estos dos supuestos se cumplen de manera satisfactoria. Los siguientes supuestos deben ser verificados.

Para verificar el cumplimiento del tercer supuesto, el de la relación lineal, usualmente se examina un gráfico de dispersión. En nuestro caso, este gráfico, tal y como es presentado por el *software* JASP, es como se ve en la figura 41.

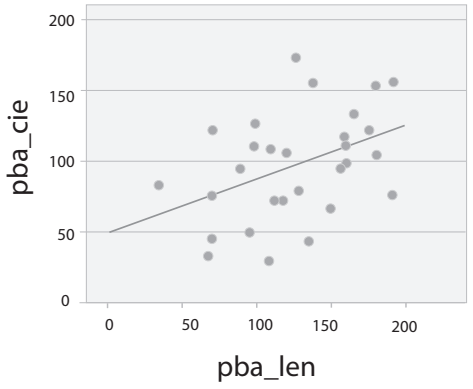


Figura 41. Diagrama de dispersión entre el puntaje en la prueba de Ciencias y en la de Lenguaje

La gráfica muestra una relación aproximadamente lineal o, lo que es más importante, no muestra una relación no lineal reconocible, hiperbólica o cuadrática, por ejemplo. Con este resultado puede asumirse el cumplimiento de este tercer supuesto.

Finalmente, para la verificación del supuesto de normalidad bivariada, el *software* JASP permite, dentro del procedimiento de correlaciones, el examen de la prueba de Shapiro-Wilk para la normalidad bivariada entre estas dos variables. El resultado de esta prueba es reportado como se muestra en la tabla 37.

Tabla 37. Resultado de la prueba de Shapiro-Wilk para la normalidad bivariada

			Shapiro-Wilk	p
pba_len	-	pba_cie	0,968	0,485

El resultado indica que el test muestra un estadístico de ,968 *p*=,485. Esta prueba asume, como hipótesis nula, la normalidad bivariada: si la distribución difiere de la curva normal, el valor de *p* sería menor que ,05. En este caso, el resultado indica que no hay diferencias entre la distribución encontrada y la curva normal; esto es, el supuesto de normalidad bivariada se verifica.

Una vez verificados los supuestos, la prueba elegida es *la prueba del coeficiente de correlación producto-momento de Pearson*. Es importante que el investigador fundamente el proceso de selección de la prueba elegida.

Si, por alguna razón, no hubiera podido asegurarse el cumplimiento de los supuestos propios de la prueba de correlación de Pearson, habríamos tenido que explorar transformaciones a las variables o, con mayor seguridad, habríamos podido elegir alguno de los equivalentes no paramétricos a la prueba de correlación de Pearson: la prueba de correlación de Spearman o la de Kendall. Estas fueron explicadas en el capítulo 4.

Paso 3. Cálculo de los estadísticos, los niveles de significación y los tamaños del efecto

Una vez que se ha seleccionado la prueba estadística adecuada y se han verificado sus supuestos propios, se dan las instrucciones al paquete de procesamiento estadístico para correr la prueba. En este punto deben explicarse conceptos fundamentales que se solicitarán al *software*: 1) el de estadístico de prueba y su nivel de significación, y 2) la potencia estadística de un experimento y el tamaño del efecto.

Estadístico de prueba, nivel de significación y tipos de error

Cuando se corre una prueba, el *software* determina, en primer lugar, el estadístico de prueba. El *estadístico de prueba* es un valor, propio de cada prueba estadística, que cuantifica la relación o la diferencia que se está evaluando. Dependiendo de la prueba, puede ser un valor *t* (para la prueba *t*), o un valor *F* (para el Anova), o un valor de chi-cuadrado (para la prueba del mismo nombre), u otros, dependiendo de la prueba. Conociendo ese valor, junto con el tamaño de la muestra, será posible calcular su nivel de significación asociado.

Para la comprensión del nivel de significación es necesario ver las consecuencias de la decisión. Para ello, conviene examinar los posibles tipos de error en las pruebas de hipótesis.

En general, en la aplicación de pruebas estadísticas se pueden presentar dos tipos de errores, conocidos como errores tipo I y II.

- *Error tipo I:* ocurre cuando se rechaza la hipótesis nula cuando esta es verdadera. En este caso se asume, de forma errónea, que la hipótesis alternativa es correcta.
- *Error tipo II*: ocurre cuando se acepta la hipótesis nula cuando esta es falsa. En este caso, la hipótesis alternativa era correcta pero la aplicación de la prueba indicó lo contrario.

En la tabla 38 se explicita de mejor manera. Las regiones sombreadas representan decisiones correctas.

Tabla 38. Decisiones correctas y tipos de error

		Condición real	
		H0 es verdadera (H1 es falsa)	**H0 es falsa (H1 es verdadera)**
Decisión	Aceptar H0	**Decisión correcta ($p = 1-\alpha$)**	Error tipo II ($p = \beta$)
	Rechazar H0	Error tipo I ($p = \alpha$)	**Decisión correcta ($p = 1-\beta$ potencia)**

Fuente: Cohen (1988).

En general, el nivel de significación es la probabilidad de cometer uno de los dos tipos de error. El nivel de significancia para el error tipo I, conocido como alfa (α), es la probabilidad de rechazar H_0 cuando esta es verdadera. Para el error tipo II, el nivel de significancia —llamado beta (β)— representa la probabilidad de no rechazar H_0 cuando esta es falsa.

En la práctica, tal vez por comodidad, se controla únicamente el error tipo I, y se minimiza la probabilidad de cometer este error, pero con frecuencia se ignora el error tipo II. Por eso se fija α (conocido como *nivel de significación* o *significancia*) en un valor pequeño, del 5 % o menos, dependiendo de las implicaciones de la decisión. Si la decisión es muy delicada, en términos médicos, por ejemplo, se fija un valor α del 1 %, o del 0,1 %, o incluso menos. En la investigación educativa y social el nivel más usual para el alfa es del 5 %.

El error tipo II se controla, en la práctica, aumentando el tamaño de la muestra. En teoría se deben calcular los dos errores; sin embargo, en la práctica esto rara vez se hace, puesto que pocos programas estadísticos calculan los dos tipos de error. Esta situación está cambiando y, dentro de las directrices del formato APA, actualmente se enfatiza en que el investigador debe dar información relacionada con los errores tipo II mediante una estimación del *tamaño del efecto*. Por esta razón, deberemos hablar brevemente de este punto.

Potencia estadística y tamaño del efecto

Definimos la *potencia estadística* de un estudio como la probabilidad de que alcance un resultado significativo si la hipótesis de investigación fuese verdadera. Esto nos permite diferenciar estudios de alta potencia, en los cuales existe una alta probabilidad de detectar como verdaderas las hipótesis realmente verdaderas, de otros con baja potencia, en los que esto puede ser difícil.

El tema de la potencia y su cálculo es complejo, por lo que no pretendemos exponerlo ampliamente en este trabajo. Tradicionalmente, se decía que la potencia del estudio debía ser asegurada de forma previa a la recolección de la información y, convencionalmente, se aceptaban estudios con, al menos, el 80 % de potencia. Definido este valor y supuesto alguna medida de tamaño del efecto esperado (alto, medio o bajo), los investigadores examinaban las tablas de potencia (p. ej., Cohen, 1988; Kraemer y Thiemann, 1987) para conocer los tamaños de muestra mínimos en cada uno de los grupos.

Este procedimiento tiene la dificultad de que, para el aseguramiento de la potencia, se requiere una idea previa del *tamaño del efecto* esperado y, como también sabemos, del tamaño de la muestra. Así, es relativamente fácil asegurar una alta potencia en un estudio simplemente aumentando el tamaño de la muestra. Otros elementos que afectan la potencia son el tipo de prueba de hipótesis seleccionada, el nivel de significación elegido y si la prueba es de una o de dos colas, si bien, la influencia de estos factores sobre la potencia es considerablemente menor.

Seguramente por estas dificultades no es frecuente que en los estudios se reporten los datos acerca de la potencia. Es presumible que con la creciente disponibilidad de *software* libre para el cálculo de la potencia esto cambie en los próximos años. Al respecto puede verse, por ejemplo, el programa G*Power, que puede ser descargado sin costo en la siguiente página web: www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3.

Este *software* puede ser realmente muy útil para el cálculo de la potencia y de los tamaños de muestra adecuados para asegurarla. Para nuestro caso, dado lo inusual de la práctica, y con el ánimo de no complejizar demasiado la exposición, no reportaremos los datos de la potencia de los estudios.

Ahora, lo que sí es actualmente parte de la práctica habitual al reportar resultados de los estudios es, además de expresar los tamaños de la muestra, indagar y expresar datos que permitan examinar *el tamaño del efecto*. Por esta razón, se dedicarán algunas líneas a su significado y a las formas de su determinación.

En general, entendemos el *tamaño del efecto* como una medida de la fuerza de un fenómeno; por ejemplo, el cambio en el resultado después de una intervención experimental. Con mucha frecuencia, después de reportar los resultados de las pruebas estadísticas y los niveles de significación alcanzados, se reportan los datos del tamaño del efecto.

Existen diferentes indicadores reconocidos para reportar el tamaño del efecto. Mencionaremos aquí varios: la llamada "*d* de Cohen" (Cohen, 1988) es la más popular, aunque no la única. Otras medidas muy populares para algunos procedimientos son el R^2 —utilizado en los modelos de regresión—, y el *eta cuadrado* (η^2), el *eta cuadrado parcial* ($\eta\rho^2$) y el *omega cuadrado* (ω^2) —usados en los análisis de varianza—. Existen muchas otras medidas de tamaño del efecto menos conocidas, entre las que se pueden mencionar los *odds ratio*, utilizados en las regresiones logísticas, la *g* de Hedges y la delta de Glass, que presentan correcciones a la *d* de Cohen, y los *common language effect sizes* (CLES), entre otras (Fritz *et al.,* 2012).

En pruebas de asociación de variables, el coeficiente de correlación producto-momento de Pearson (*r*) es, en sí mismo, una medida de tamaño del efecto. Existen otras medidas de asociación que se utilizan para calcular el tamaño del efecto en algunas pruebas específicas: la correlación rango biserial (r_b) —usada para examinar asociaciones entre variables dicotómicas y variables ordinales— o el coeficiente *V de Cramer* —empleado para examinar asociaciones entre variables nominales— son ejemplos de ellas.

Conociendo algunas medidas, es posible calcular otras. La tabla 39 presenta la interpretación de algunas de las medidas más populares de tamaño del efecto.

Tabla 39. Interpretación de algunas medidas de tamaño del efecto

***d* de Cohen**	**r, r_b**	**η^2**	**h_p^2, w^2**	**Interpretación según Cohen (1988)**
< 0	< 0	-	-	Efecto adverso
0,0	.00	0,00	0,00	No hay efecto
0,2	.10	0,01	0,01	Efecto pequeño
0,5	.30	0,09	0,06	Efecto mediano
0,8<	.50<	0,25<	0,14<	Efecto grande

Debido a las dificultades que presentan algunos programas, tales como el IBM-SPSS en las versiones menos actualizadas, para el cálculo del tamaño del efecto se sugiere a los investigadores que prefieran este tipo de *software* el uso de calculadoras, programas o páginas web dedicadas al cálculo del tamaño del efecto en pruebas específicas. Una de las páginas web más útiles para ello es la página alemana *Psychometrica*, que se encuentra en la siguiente dirección: https://www.psychometrica.de/effect_size.html.

Este punto es muy importante. Actualmente, el manual de publicaciones de la APA, desde la versión 6, requiere que, junto con los estadísticos, los grados de libertad y los niveles de significación de la prueba, se aporte alguna medida de tamaño del efecto y su interpretación. En lo que sigue, se ilustra cómo se haría esto en nuestro ejemplo.

Ejemplo (continuación). Cálculo de estadísticos, niveles de significación y tamaños del efecto

Recuadro 22. Ejemplo (continuación). Relaciones entre dos pruebas de logro

3. Cálculo de los estadísticos, niveles de significación y tamaños del efecto

En nuestro ejemplo el tamaño de la muestra es de treinta alumnos de grado 10 de una institución educativa. El nivel de significación aceptado lo fijaremos, como es tradicional, en α =,05.

La salida del *software* JASP para la correlación entre las dos pruebas, especificando que se reporte por parejas de variables (*display pairwaise*) y que se reporten los límites del intervalo de confianza, es como sigue:

Tabla 40. Correlación producto-momento de Pearson entre las pruebas de Lenguaje y Ciencias

Correlaciones de Pearson					
			***r* de Pearson**	**p**	**Intervalo de confianza inferior al 95%**	**Intervalo de confianza superior al 95%**
pba_len	-	pba_cie	0,423	0,020	0,074	0,680

De acuerdo con los resultados, el valor de la correlación producto-momento de Pearson para estas dos variables es de *r*=,423 y la probabilidad (bilateral) de cometer el error tipo I en esta situación, es decir, de rechazar H_0 cuando no existen diferencias, es de *p*=,020. Este valor es menor que ,05, que habíamos fijado previamente, lo cual da evidencia empírica suficiente para rechazar la hipótesis nula y concluir que sí existe una asociación significativa entre los resultados de las dos pruebas.

Es interesante notar que este valor de probabilidad se presenta en la medida en que habíamos especificado un nivel de probabilidad bilateral en el *software* que la calcula. Si especificamos una probabilidad unilateral, más acorde a la forma en que habíamos definido originalmente la hipótesis, esta probabilidad baja aún más, hasta *p*=,010, exactamente la mitad de la bilateral.

En el extremo derecho de la tabla, el *software* ha añadido dos datos nuevos que corresponden a los límites, superior e inferior, del intervalo de confianza (CI: *confidence interval*), al 95 %, para la correlación de Pearson. En términos sintéticos estos datos nos dicen que, con un 95 % de confianza, la verdadera correlación *r* de Pearson se encuentra en el intervalo [0,074, 0,680]; o sea, es mayor que 0,074 y menor que 0,680. El significado de este intervalo de confianza será examinado más adelante, en este mismo capítulo.

Una vez hemos aplicado la prueba y examinado el nivel de significación alcanzado, debemos proceder al examen del tamaño del efecto. Para el caso del coeficiente de correlación producto-momento de Pearson debe anotarse que, en sí mismo, este coeficiente puede ser entendido como una estimación del tamaño del efecto, en la medida en que representa la magnitud de la asociación. Otra posibilidad de mostrar aquí una medida de tamaño del efecto es presentar la proporción de varianza explicada, o compartida entre las dos variables, esto es, el cuadrado del coeficiente de correlación (r^2), que en este caso es $r^2=\eta^2=,179$.

En la medida en que, en sí mismos, estos valores representan estimaciones válidas de tamaño del efecto, no se acostumbra a aportar nuevas estimaciones para las correlaciones de Pearson. Para el caso de la estimación del tamaño del efecto de una correlación de Pearson, preferiríamos guiarnos por la nota de Cohen que señala que un valor *r* entre ,3 y ,5 debe ser considerado como un tamaño intermedio del efecto.

Paso 4. Reporte de los resultados

Formato para reportes de resultados en texto

Después de aplicar las pruebas se interpretan los resultados y se reportan en un formato adecuado. Al respecto de la pregunta de si la correlación entre los dos puntajes es estadísticamente significativa, debemos observar que, en general, la correlación entre dos variables siempre va a ser diferente de cero, simplemente por el hecho de que trabajamos con muestras de la población. Sin embargo, es posible diferenciar entre correlaciones aleatorias y correlaciones debidas a asociaciones efectivamente presentes en la totalidad de la población. Así, la respuesta a esta pregunta se da en términos de la probabilidad de cometer el error tipo I.

En general, para reportar los resultados de las pruebas de hipótesis, existe un cierto formato que indica que, después de expresar verbalmente las pruebas y su interpretación, deben anotarse los resultados numéricos del estadístico, su nivel de significación y, cuando procede, una medida de tamaño del efecto y de los intervalos de confianza.

Esto, en nuestro ejemplo, podría ser expresado, en texto, utilizando el siguiente formato:

r=<valor de r> *p*=<valor de p>

Si se va a reportar el intervalo de confianza del coeficiente de correlación, el formato será el que sigue:

r=<valor de *r*> *p*=<valor de *p*> IC <intervalo>% [<límite inferior>, <límite superior>]

Para el caso de nuestro ejemplo, los resultados indicaron que existe una correlación estadísticamente significativa que, al tiempo, se corresponde con un tamaño del efecto intermedio.

Ejemplo (continuación): reporte de resultados

Recuadro 23. Ejemplo Relaciones entre dos pruebas de logro (continuación).

5. Reporte de resultados

Los resultados del examen de la prueba de hipótesis sobre la relación entre la prueba de Lenguaje y la de Ciencias Naturales indicaron una correlación producto-momento de Pearson positiva y significativa al nivel de ,05, que se corresponde con un tamaño de efecto intermedio *r*=,42 *p*=,023 IC 95 % [,08, ,68].

Una alternativa a la prueba de hipótesis: la estimación y los intervalos de confianza

La prueba de hipótesis es el tema principal del presente libro. Existe, sin embargo, una forma alternativa o complementaria a la prueba de hipótesis, relacionada con la estimación de parámetros poblacionales a partir de los estadísticos obtenidos en una muestra.

La mejor estimación de un parámetro poblacional es el estadístico correspondiente en la muestra. En otras palabras, la mejor estimación de la media poblacional será la media muestral; de igual forma, la mejor estimación de la desviación estándar poblacional será la desviación estándar muestral, igual con las correlaciones entre dos variables o con cualquier otro parámetro poblacional.

Ahora: pueden distinguirse dos formas de estimar estos parámetros. La primera es, simplemente, asumir el estadístico correspondiente, obtenido en la muestra; a esto lo llamamos *estimación puntual*. La segunda es encontrar un intervalo verosímil alrededor del estadístico muestral en el que podemos asegurar que se encuentra ese parámetro. Esto se conoce como una *estimación por intervalos*.

Cuando se hace una estimación por intervalos, se determina un nivel de confianza (usualmente de 95 %, aunque ocasionalmente se encuentra también del 99 %) que representa la seguridad que tenemos de que el parámetro poblacional estará en ese intervalo. Esto se conoce como *intervalo de confianza* (IC). En otras palabras, cuando enunciamos un intervalo del 95 % de confianza, realmente estamos diciendo que estamos, en un 95 % seguros, de que el parámetro poblacional se encuentra dentro de nuestro intervalo.

Para el cálculo del intervalo de confianza, ya antes hemos introducido un concepto importante: el del *error estándar*. Con este concepto es fácil definir el intervalo de confianza utilizando el teorema de Chevichev. El intervalo de confianza del estimador al 95 % es, aproximadamente, el estimador más o menos dos veces el error estándar. Al 99 % será el estimador más o menos tres veces el error estándar.

En este caso, como en otros que nos preceden, no nos preocuparemos por los detalles del cálculo del intervalo de confianza y de los límites de confianza de los diferentes estadísticos. Baste decir que puede plantearse una clara relación entre los intervalos de confianza y la prueba de hipótesis,

de la siguiente forma: si un intervalo de confianza no incluye el cumplimiento de la hipótesis nula, entonces, el resultado es estadísticamente significativo.

En el caso de nuestro ejemplo, el intervalo de confianza al 95 % para la correlación de Pearson mostró un límite inferior de 0,074 y un límite superior de 0,680. Esto señala que el IC, al 95 %, fue de [0,076, 0,680]. En la medida en que el valor cero (0) para la correlación de Pearson no quedó incluido en el IC al 95 %, podemos afirmar que el resultado es significativo (con un 95 % de seguridad).

Este punto nos vincula la estimación de parámetros por intervalos de confianza con las pruebas de hipótesis. Al respecto, algunos investigadores han llegado a proponer que los intervalos de confianza podrían sustituir las pruebas de hipótesis (p. ej., Cohen, 1994; Hunter, 1997; Schmidt, 1996, citado por Aron y Aron, 2001). Los que defienden esta posición se fundamentan en la idea de que los intervalos de confianza dan toda la información importante de una prueba de significación, e incluso aportan más que la presentada por la prueba de hipótesis.

A pesar de estas posiciones, los intervalos de confianza no tienen aún un uso muy generalizado en las publicaciones científicas, aunque cada vez es más fácil encontrarlos, especialmente al contrastar hipótesis de diferencias de medias entre dos muestras independientes o al estimar coeficientes de regresión.

Creemos que la mejor recomendación al respecto es incluirlos como información complementaria a la prueba de hipótesis. En la siguiente sección se mostrará cómo deben ser presentados en publicaciones científicas.

Reporte de los resultados de pruebas en publicaciones científicas

Aunque todo el proceso de la prueba de hipótesis se basa en la presencia de una hipótesis nula, que se pretende desestimar, y de una hipótesis alternativa que se pretende verificar, estas hipótesis nunca aparecen en las publicaciones científicas, ya que se supone que el lector comprende perfectamente este proceso.

En general, cuando se escriben los resultados de investigación se indica, en primer lugar, la prueba que se utilizó, si esta arrojó resultados significativos (utilizando para ello un formato que resulta propio de cada prueba) y, a continuación, se presentan una o varias medidas de tamaño del efecto y, de estar disponibles, los intervalos de confianza correspondientes.

Cuatro convenciones, utilizadas en formato APA, que deben ser tenidas en cuenta a la hora de presentar resultados estadísticos son las siguientes:

- Convención 1. Número de decimales. En general, los diferentes indicadores, parámetros y estadísticos deben ser presentados con dos decimales. Excepción a esta regla es el valor de la significación alcanzada, que debe ser siempre presentada con tres decimales (APA, 2010, p. 115).
- Convención 2. Cuando un indicador no puede superar el valor de 1, como es el caso del valor *p*, o de los coeficientes de correlación de Pearson, Spearman o Kendall, por ejemplo, se omite la escritura del número cero, en la posición de los enteros. Esto significa que, si la prueba mostró un nivel de significación de 0,023, este se escribe como p = ,023; si el coeficiente de correlación producto-momento de Pearson es de 0,85, esto se comunicará de la forma "r = ,85" (APA, 2010, p. 114).

- Convención 3. Cuando se expresa el intervalo de confianza para un valor, debe expresarse la estimación puntual del valor y el intervalo de confianza siguiendo este formato

  IC x % [LI, LS]

  en donde "x" representa del porcentaje del intervalo de confianza (usualmente 95 % o 99 %), "LI" el límite inferior del intervalo y LS el límite superior del mismo.

  Por ejemplo, para reportar el intervalo de confianza de una correlación de Pearson en texto, se expresaría : *r*=,85 IC 95 % [,81, ,89]

- Convención 4. Con mucha frecuencia, los artículos presentan los resultados de muchas pruebas de hipótesis en tablas, cada una con sus correspondientes niveles de significación, tamaños del efecto e intervalos de confianza. En estos casos, los niveles de significación son frecuentemente acompañados con sucesiones de asteriscos (*, **, ***) que representan los valores de las significaciones alcanzados, en los niveles convencionalmente aceptados (esto es, en menos del 5 %, o en menos del 1 %..). Esto se expresa como se muestra en la tabla 41.

Tabla 41. Tabla de convenciones comúnmente usadas para indicar niveles de significación alcanzados

NS	No significativo
*	,01<p<,05
**	,001<p<,01
***	p<,001

Esta convención permite apreciar, con un golpe de vista, los niveles de significación y compararlos fácilmente. Por supuesto, es necesaria una nota haciendo explícita esta convención en cada tabla que la utilice.

Capítulo 9

Las pruebas estadísticas, supuestos y transformaciones

Distribuciones muestrales de probabilidad

Una *distribución muestral de probabilidad* es la distribución de probabilidad de un estadístico que se construye con todas las muestras aleatorias posibles de tamaño *n*. La distribución conjunta de los resultados se puede ajustar a una función de probabilidad con todas sus propiedades, valor esperado y varianza. Con esta distribución se puede calcular la probabilidad de ocurrencia de un estadístico para *cualquier* tamaño de muestra.

Tradicionalmente, se publicaban extensas tablas en las que, para cada tamaño de muestra y para cada valor del estadístico, se presentaban los valores de probabilidad asociados. Con estas tablas, los investigadores determinaban los niveles de significación en las pruebas de hipótesis. Hoy en día, los paquetes estadísticos calculan directamente el estadístico y la probabilidad asociada con este para un tamaño de muestra dado.

Como ya lo hemos mencionado, las pruebas de hipótesis se clasifican en paramétricas y no paramétricas. En las primeras, se asume que los datos de la muestra provienen de una población modelada por una distribución de probabilidad conocida. Por esta razón, es necesario validar estos supuestos de distribución, especialmente el de la distribución normal (Z), de forma previa al cálculo de la probabilidad del estadístico de prueba.

Existen varias relaciones entre las distribuciones muestrales de probabilidad, por ejemplo, la distribución *t*, que es una distribución muestral, se utiliza en lugar de la distribución Z, porque para muestras de menos de treinta observaciones, la distribución *t* es más confiable y para muestras mayores, los resultados son prácticamente iguales.

Otra distribución paramétrica muy conocida es la F. El análisis de varianza (Anova), utilizado para comparar tres o más medias, maneja la distribución muestral F para calcular la probabilidad de igualdad de las medias. Como en el caso de la prueba *t*, se conocen las propiedades y valores de la distribución muestral de F para diferentes tamaños de muestra.

Otra distribución muestral conocida y ampliamente utilizada es la Chi-cuadrado. Esta es una distribución muestral continua, pero se utiliza para evaluar pruebas no paramétricas, como la prueba

de independencia de variables categóricas, que no necesita supuestos de distribución normal. La distribución Chi-cuadrado cambia según la cantidad de filas y columnas que tenga la tabla de cruce de las dos variables (más adelante esto será relacionado con el concepto de grados de libertad).

Por otro lado, existe una relación entre las pruebas no paramétricas y distribución normal. Para muestras grandes, varios de los estadísticos de prueba convergen a la distribución normal estándar, por lo que los paquetes estadísticos suelen utilizar la distribución Z (o normal estándar) para el cálculo de la significación de estos estadísticos.

La elección de la prueba estadística: una visión general de las pruebas

En términos generales, las pruebas pueden diferenciarse dependiendo de si se utilizan para 1) examinar asociaciones o correlaciones o 2) contrastar muestras.

El primer caso ya lo hemos examinado brevemente, en el tema de la estadística descriptiva, cuando presentamos los coeficientes de correlación de Pearson, Spearman y Kendall, así como los coeficientes de asociación entre variables nominales (coeficiente de contingencia *C*, *V* de Cramer, Phi). De igual forma, desarrollamos nuestro ejemplo de las pruebas de hipótesis utilizando la correlación de Pearson. La selección de la prueba estadística dependerá acá, básicamente, de tres puntos: 1) el nivel de medición de las variables, 2) el tipo de asociaciones presentes (lineales/no lineales) y 3) el número de variables implicadas (dos, tres o más de tres).

El segundo caso, en el que nos interesa contrastar muestras, es bastante más extenso y mucho más complejo. En principio, podemos clasificar las pruebas si se utilizan para 1) contrastar una medida con un valor, 2) contrastar dos medidas o 3) contrastar tres medidas o más. En el primer caso, las pruebas que contrastan una muestra lo hacen con un valor, y se diferencian dependiendo del nivel de medida de la variable. El segundo y el tercer caso, en donde se pretenden contrastar los resultados de dos o tres muestras, requieren diferenciar si estas muestras son independientes o "apareadas" y, a partir de allí, el nivel de medición de las variables y el cumplimiento de los supuestos determinará la prueba elegida.

El siguiente esquema presenta esta clasificación de las diferentes pruebas. Aquí se señalan las pruebas que se tratan en este trabajo con un asterisco (*), y aquellas que se incluyen en el SPSS, con una "s" o en JASP, con una "j".

1. Interesa examinar asociaciones o correlaciones.

- Intervalo: examinar la linealidad de la relación con una gráfica de dispersión y con una prueba de linealidad.
  - Las relaciones son lineales*[s].
    - Dos variables: coeficiente de correlación de Pearson*[sj]
    - Tres o más variables: correlación múltiple y parcial[sj]
  - Las relaciones no son lineales: examinar modelos de regresión no lineal[s].

- Ordinales:
  - Dos variables: correlación de Spearman*sj, correlación de Kendall*sj
  - Tres o más variables: coeficiente de concordancia de Kendallsj
- Variables nominales:
  - Medidas simétricas: coeficiente de contingencia C*sj, coeficiente *V* de Cramer*sj, coeficiente Phi*sj
  - Medidas direccionales: coeficiente Lamda*s, coeficiente Tau de Grossmann y Kruskal*s y coeficiente de incertidumbre*s

2. Interesa contrastar muestras.

- Una muestra (una variable contra un valor constante)
  - Intervalo: prueba de valor Z, prueba *t* de una muestrasj
  - Ordinal: prueba de Kolmogorov-Smirnovsj
  - Variable nominal: prueba binomial, Chi de una muestrasj
- Dos muestras
  - Independientes (la misma variable en dos muestras separadas).
    - Intervalo:
      - Con homogeneidad de varianzas: *t* de Student*sj.
      - Sin homogeneidad de varianzas: prueba *t* de Student-Welch*sj
    - Ordinal: prueba U de Mann-Whitney*sj
    - Nominal
      - Probabilidad exacta de Fischer (n<20) sj
      - Chi-cuadrado (n>20) *sj
  - Apareadas (dos variables en la misma muestra)
    - Intervalo: *t* de Student para muestras dependientes*sj
    - Ordinal: Prueba de Wilcoxon*sj
    - Nominal:
      - Dicotómica: prueba de McNemar*s
      - Politómica: prueba de McNemar-Bowker*s
- Tres o más muestras
  - Independientes
    - Intervalo:
      - Con homogeneidad de varianzas:
        - » Análisis de varianza de una entrada Anova (sin buscar interacción)*sj
        - » Análisis de varianza de doble entrada (buscando interacción de factores)sj
      - Sin homogeneidad: Prueba de Brown-Forsythe*sj, Prueba de Welch*sj

- Ordinal: prueba H de Kruskal-Wallis*[sj]
- Nominal
  - Chi de Pearson (todas las casillas con n>5)*[sj]
  - Chi de proporciones (alguna casilla con n<5)

– Apareadas

- Intervalo: análisis de varianza de medidas repetidas (Anova MR)*[sj]
- Ordinal: análisis de varianza de doble entrada de Friedman*[sj]
- Nominal: prueba Q de Cochran*[s]

Los supuestos de las pruebas y su verificación

En las pruebas de hipótesis de tipo paramétrico o en la construcción de modelos estadísticos, es necesario verificar los supuestos para poder utilizar el método, la prueba o el modelo estadístico. Los supuestos más comunes e importantes son:

- *Supuesto de normalidad.* Los supuestos de normalidad son necesarios para utilizar estadísticos de prueba como la prueba *t* o el análisis de varianza, dado que estas distribuciones muestrales se construyeron a partir de la distribución normal estándar. Usualmente, se formulan en términos de que una variable determinada debe distribuirse de forma normal o "aproximadamente normal".
- *Supuesto de homogeneidad de varianzas (homocedasticidad o igualdad de varianzas).* Este supuesto es importante cuando la prueba se basa en la variación de los datos, como análisis de varianza**.** Se formula en términos de que las varianzas de una determinada variable deben ser similares para todos los valores de otra variable.
- *Supuesto de independencia.* Este supuesto se utiliza principalmente en la construcción de modelos estadísticos tales como el modelo de regresión. El supuesto consiste en que las variables independientes del modelo deben ser efectivamente independientes entre sí. Esta independencia se verifica en el análisis de los residuos del modelo cuando se encuentra que los residuos se distribuyen de forma normal con media cero y varianza constante.

Para la verificación de los supuestos, se examina a través de pruebas específicas que determinan, directamente, si el supuesto se cumple o no. En lo que sigue, examinaremos las diferentes pruebas para verificar el cumplimiento de estos supuestos. En el caso en que la prueba indica que el supuesto no se cumple, es posible proceder a través procesos de transformación de los datos, haciendo que nos aproximemos de mejor medida al cumplimiento del supuesto.

Es importante anotar que, en algunas situaciones, no es posible lograr que un conjunto de datos cumpla todos los supuestos que se le exigen para el uso de una prueba particular. En estos casos es importante comprender el efecto que pueda tener la violación de ese supuesto en esa prueba, en particular. Algunas pruebas no se ven gravemente afectadas por la violación de algún supuesto; en el argot, se dice que estas pruebas son "robustas" frente a la violación del supuesto. En otras, ciertas violaciones invalidan totalmente el uso de la prueba. En algunas pruebas, algunas violaciones involucran, por ejemplo, una mayor probabilidad de incurrir en un error de tipo I. Al respecto, es importante conocer cada prueba.

Pruebas para la verificación de supuestos

Normalidad: pruebas de Kolmogorov-Smirnov y Shapiro-Wilk

En casi todos los análisis estadísticos paramétricos se supone que los datos recolectados tienen una distribución normal. Este contraste se realiza previamente de tal forma que los análisis posteriores sean fiables.

La *prueba de Kolmogorov-Smirnov*, llamada prueba K-S, se emplea preferentemente con variables numéricas de intervalo o razón, para comprobar que los datos siguen una distribución normal. Contrariamente a lo que se desea en la mayoría de los casos, en las pruebas de normalidad se busca aceptar la hipótesis nula (H_0). Así, el valor $p \geq 0{,}05$ en los test de normalidad indicaría que no hay evidencia suficiente para rechazar la normalidad de la variable. La prueba K-S es muy utilizada con muestras mayores a cincuenta observaciones.

Las hipótesis de esta prueba se formulan como sigue:

- Hipótesis nula (H_0). El conjunto de datos sigue una distribución normal.
- Hipótesis alternativa (H_1). El conjunto de datos no sigue una distribución normal.

Para correr la prueba de Kolmogorov-Smirnov sobre una variable, en el SPSS, puede seguirse el este camino:

> */Analizar/Estadísticos descriptivos/Explorar… en este punto, se selecciona la variable que se desea examinar en la "lista de dependientes" y, en el menú desplegado en el botón "Gráficos", se selecciona la casilla "Gráficos con pruebas de normalidad".*

Ejemplo. Se tiene una muestra de 247 estudiantes a quienes se les aplicó un test para detectar su tendencia a mostrar un patrón de aprendizaje de tipo MD, o "dirigido al significado". La escala tiene un mínimo posible de 1 punto y un máximo posible de 5 puntos. La gráfica de la figura 42 muestra el histograma de esta variable.

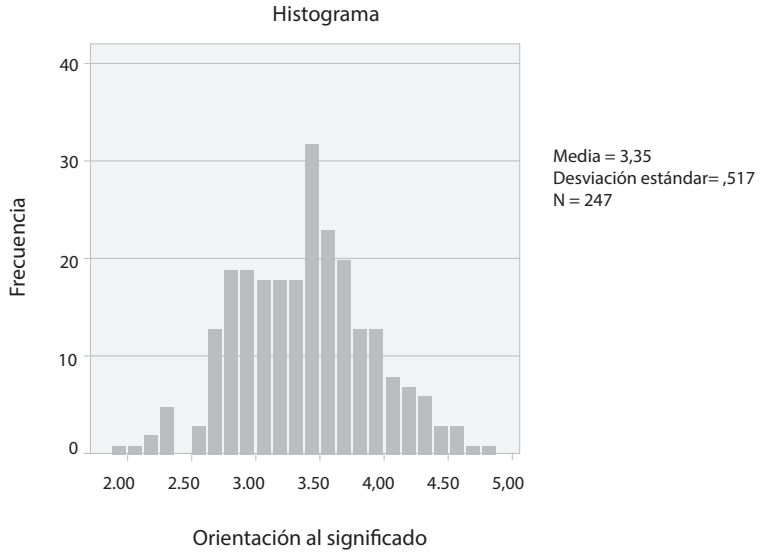


Figura 42. Histograma de la variable "orientación al significado"

La tabla 42 muestra los resultados arrojados por el SPSS. De acuerdo con los resultados, la prueba indica que el estadístico de Kolmogorov-Smirnov es de K-S(247) = 0,048 p=,200. En la medida en que el nivel de significación obtenido es mayor que ,05, asumimos que la distribución no difiere de forma significativa de la distribución normal.

Tabla 42. Resultados de las pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**	**Estadístico**	**gl**	**Sig.**
MD **Orientación al significado**	,048	247	,200*	,994	247	,436

***Esto** es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors.

En esta misma tabla, el SPSS presenta los resultados de otra prueba frecuentemente usada para la validación del supuesto de normalidad: la *prueba de Shapiro-Wilk*. En general, se considera más apropiado el uso de la prueba de Shapiro-Wilk cuando se tienen muestras menores a cincuenta individuos mientras que el uso de la prueba de Kolmogorov-Smirnov se considera más adecuado para muestras de más de cincuenta.

Este procedimiento arroja también una forma visual de examinar la normalidad con la gráfica P-P. En esta se distribuyen los datos de la variable de menor a mayor, y cuanto más se ajusten los datos a la diagonal más cercana es la variable a la distribución normal (figura 43). Estos resultados son coherentes con los resultados de la prueba K-S.

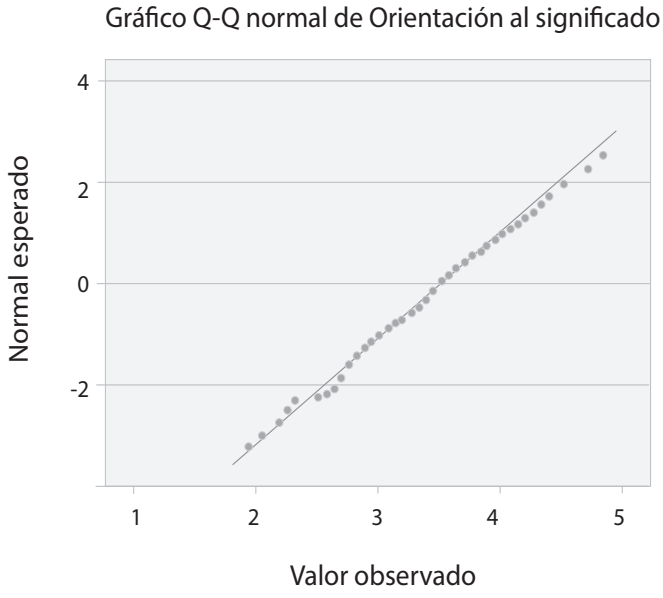


Figura 43. Gráfica Q-Q normal de la variable "orientación al significado"

En los casos en que no se verifica el supuesto de normalidad, puede procederse a una transformación de los datos para intentar que la variable cumpla este supuesto o, al menos, se aproxime de mejor forma a este cumplimiento. Al final de este capítulo se expondrán algunas formas de transformar los datos.

Homocedasticidad: la prueba de Levene

Cuando existen varios grupos de sujetos y queremos examinar diferencias en las medias de una o más variables, es necesario verificar el supuesto de *homocedasticidad*, es decir, que las varianzas de la variable en los diferentes grupos sean estadísticamente iguales.

Iniciando con el caso en el que queremos comparar las medias de una variable métrica entre dos o más grupos, la prueba más utilizada para verificar este supuesto es el llamado *test de Levene.* Para el test, las hipótesis se plantean de la siguiente forma:

- Hipótesis nula (H_0). La variabilidad de los grupos es homogénea.
- Hipótesis alternativa (H_1). La variabilidad de los grupos es diferente.

Como se observa, al igual que para los test de normalidad, en el caso del test de Levene esperamos aceptar la hipótesis nula.

En general, la prueba de Levene se presenta como opción en los diferentes procedimientos que la requieren como supuesto, tales como la prueba *t* de Student para grupos independientes (en este caso, el SPSS calcula automáticamente la prueba de Levene), el análisis de varianza en una dirección (Anova de una vía) o el análisis de varianza de medidas repetidas (Anova MR). Todas estas pruebas de hipótesis serán examinadas en detalle más adelante.

Existen algunas pruebas específicas que presentan variaciones del supuesto de homogeneidad de varianzas. Este es el caso, por ejemplo, del análisis de varianza de medidas repetidas (Anova MR), utilizado para el examen de diferencias entre varias tomas de la misma variable en la misma muestra de población. Para esta prueba, uno de los supuestos es el de la igualdad de las varianzas de todas las posibles diferencias entre las mediciones; lo que se conoce como "supuesto de esfericidad", y se evalúa a través de la *prueba* Ω *de Mauchly*, la cual se examinará con más detalle cuando presentemos el Anova MR.

Ahora, si en el estudio van a ser objeto de análisis simultáneo dos o más variables dependientes, para la comparación de la igualdad de las matrices de varianza/covarianza, se utiliza el llamado *test M de Box*. Este test es utilizado en estadística multivariante en procedimientos que lo requieren como supuesto, tales como el análisis multivariante de varianza (Manova), que no será tratado en este trabajo.

Si se desea examinar la homogeneidad de varianzas entre grupos de forma independiente de los procedimientos que la requieren como supuesto, es posible hacerlo en el SPSS a través del procedimiento "Explorar". Para hacerlo, puede seguirse el camino presentado en el recuadro 24.

Recuadro 24. Para examinar la homogeneidad de varianza en IBM-SPSS

/Analizar/Estadísticos descriptivos/Explorar … en este punto, se selecciona la variable que se desea examinar en la "lista de dependientes" y una variable nominal (factor) en la "lista de factores". En el menú desplegado en el botón "Gráficos", debe haberse activado la opción "Dispersión versus nivel con prueba de Levene"; se selecciona allí la casilla "No transformados".

Cuando este procedimiento se sigue en la variable que examinamos anteriormente relacionada con el patrón de aprendizaje MD o "dirigido al significado", con respecto a la variable género, el SPSS arroja la tabla 43.

Tabla 43. Prueba de homogeneidad de varianza

		Estadístico de Levene	gl1	gl2	Sig.
MD Orientación al significado	Se basa en la media	,956	1	245	,329
	Se basa en la mediana	,913	1	245	,340
	Se basa en la mediana y con gl ajustado	,913	1	235,812	,340
	Se basa en la media recortada	,938	1	245	,334

Tomamos de la tabla la prueba basada en la media. De acuerdo con los resultados, el estadístico de Levene es de F(1)= 0,956 p =,329 > ,05. Este resultado indica que las varianzas son iguales entre los dos grupos de género. Hay homocedasticidad, por tanto, este supuesto, requerido para una prueba *t*, se cumple.

La figura 44, de cajas y bigotes, ilustra el cumplimiento de este supuesto. Como se observa, la dispersión de la variable, indicada por la longitud de las cajas, es bastante similar en los dos grupos de género.

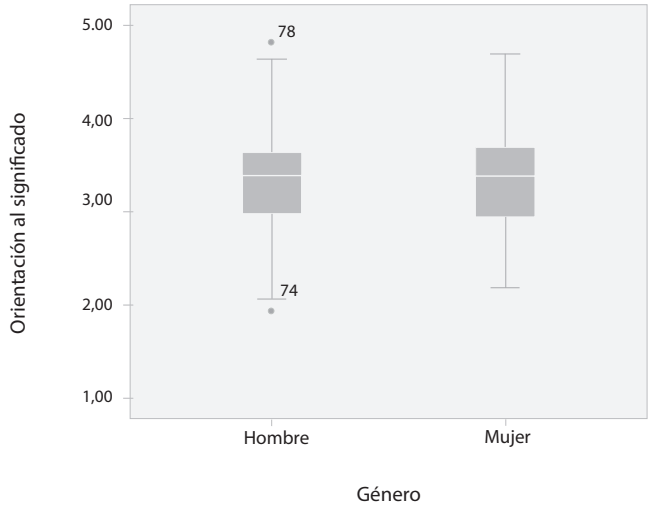
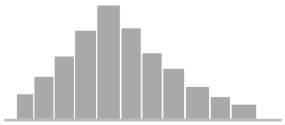
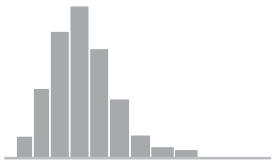
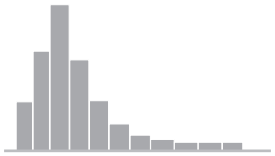
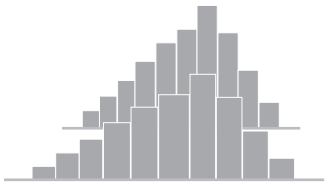


Figura 44. Gráficas de cajas y bigotes de la variable "orientación al significado" de forma separada para grupos de género

Transformaciones de los datos

En algunas ocasiones, la falta de normalidad de una variable puede subsanarse mediante una transformación de esta. A este procedimiento se le denomina *transformación de datos*. Una vez se haya realizado dicha transformación, si se cumplen los demás supuestos estadísticos, se podrá calcular el procedimiento que lo requería como supuesto. En la tabla 44 se muestran algunas de las transformaciones más utilizadas, dependiendo de la forma en que ha violado el supuesto de normalidad.

Tabla 44. Transformaciones más utilizadas en distribuciones que violan el supuesto de normalidad

Forma	Forma de la distribución	Representación gráfica	Transformación aconsejada
Asimetría positiva	Moderadamente asimétrica hacia la derecha		**√X**
	Marcadamente asimétrica hacia la derecha		**Log(X+C)**
	Extremadamente asimétrica hacia la derecha (Leptocurtosis)		**1/X**
Asimetría negativa	Asimetría negativa		**Log(C-X)**
	Platicurtosis		**X^2**

Para distribuciones asimétricas positivas se usan las transformaciones $\sqrt{x}$, log(x) y 1/x, las cuales comprimen los valores altos y expanden los pequeños. El efecto de estas transformaciones está en orden creciente: menos efecto $\frac{1}{\sqrt{x}}$, más efecto log(x) y un efecto mucho mayor es 1/x.

Cuando se tienen distribuciones de frecuencias con asimetría negativa (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar la transformación $y = x^2$ o y = log(C-X). Estas transformaciones comprimen la escala para valores pequeños y la expanden para valores altos.

Por ejemplo, se tienen las calificaciones en Ciencias Naturales de 22 estudiantes de una institución educativa de Bogotá; la máxima calificación es 100. La prueba de normalidad Shapiro-Wilk, apropiada para esta situación, arroja los resultados presentados en la figura 45.

Pruebas de normalidad						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Calificación en Ciencias Naturales	,282	22	,000	,646	22	,000

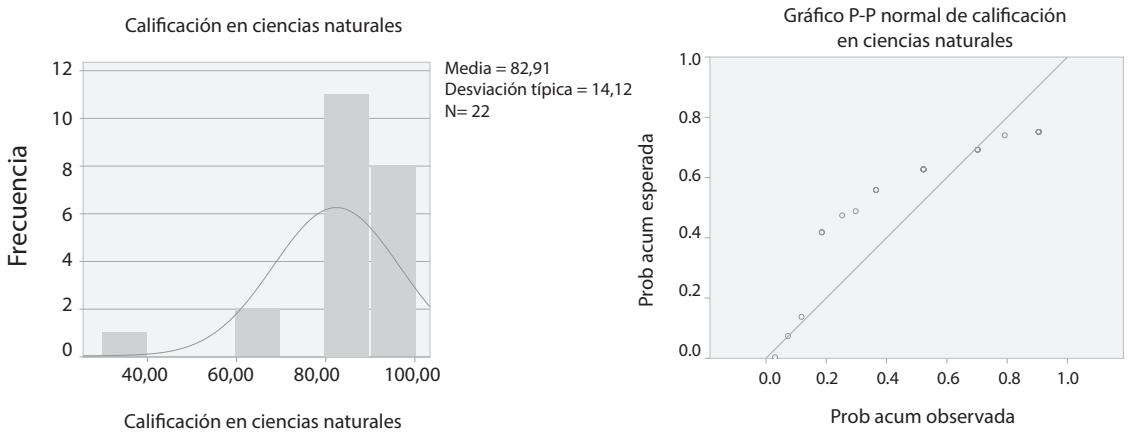


Figura 45. Prueba de normalidad, histograma y gráfico P-P de la variable "calificación en ciencias naturales"

Los resultados obtenidos ponen de manifiesto la falta de normalidad de la variable "calificaciones en Ciencias Naturales", siendo la razón su elevado grado de asimetría negativa.

Una posible solución a la falta de normalidad de esta variable es transformarla logarítmicamente. En este caso la transformación a aplicar sería Log(C-X). Específicamente:

Y=Log(95-calificación en Ciencias Naturales)

La constante (C), tiene que ser un número tal que permita que las observaciones no tomen números negativos. Para este caso, en particular, a la constante C se le ha asignado el número 95.

En la figura 46 se muestran los resultados obtenidos al analizar la normalidad de dicha variable. De acuerdo con los datos, no se observan desviaciones significativas de la hipótesis de normalidad, como lo indica la prueba de Shapiro-Wilk W(22)=0,939 *p*=,192. Puesto que el valor del nivel crítico es mayor que 0,05, asumimos que esta distribución no difiere significativamente de una distribución normal.

Pruebas de normalidad						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Calificación transformada	,120	22	,200*	,939	22	,192

a. Corrección de la significación de Lilliefors.
*Este es un límite inferior de la significación verdadera.

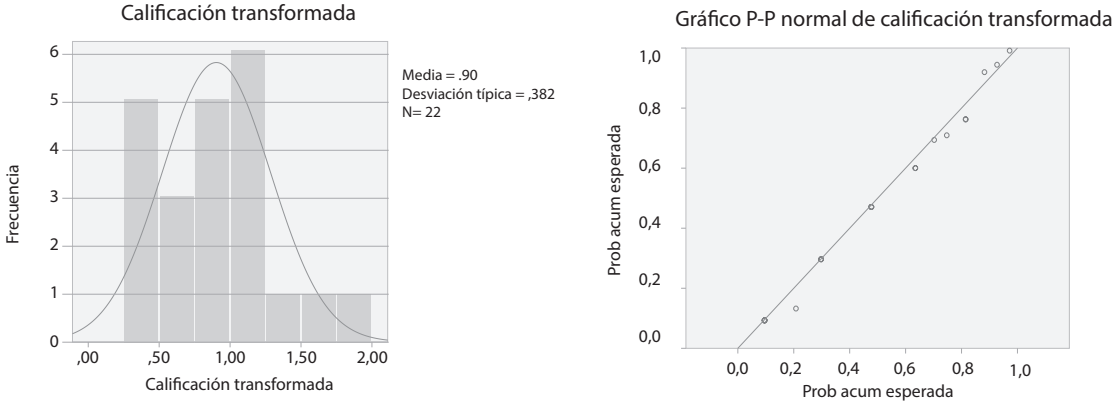


Figura 46. Prueba de normalidad, histograma y gráfico P-P de la variable "calificación en ciencias naturales transformada"

Formas en que se expresan los supuestos en publicaciones científicas

Los supuestos que hemos examinado se verifican a través de la aplicación de pruebas específicas que, en su mayoría, tienen como hipótesis nula la condición de cumplimiento del supuesto. Para reportar los resultados de una prueba en texto, esta debe ser expresada y, en general, se debe seguir un formato que, aproximadamente, es así:

<Nombre de Estadístico>(<gl>) = <Valor de estadístico> p =/< <valor de p>

Para el caso de la variable transformada del ejemplo anterior, esto podría expresarse de la siguiente forma:

> *Los resultados de la prueba de Shapiro-Wilk indicaron que la variable original, calificación en Ciencias Naturales, difiere de forma significativa de la curva normal W(22)=0,65 p<,001, por lo que se procedió a la aplicación de la transformación Log(95-calificación en Ciencias Naturales). Los resultados de la aplicación de prueba de Shapiro-Wilk sobre la variable transformada mostraron que su distribución no difiere de forma significativa de la curva normal W(22)=0,94 p=,192.*

Capítulo 10

Pruebas de diferencias entre dos medidas

Una buena parte de los diseños de investigación educativa y social requieren de la determinación de la diferencia entre dos observaciones.

Tenemos que determinar diferencias cuando comparamos los resultados de dos subgrupos presentes en una muestra, hombres y mujeres, por ejemplo, o estudiantes que recibieron un tratamiento educativo frente a otros que no lo recibieron. Este tipo de diseños, en los que tenemos una medida y comparamos los resultados de dos subgrupos en dicha medida se conocen como diseños *intersujeto*. En ellos, comparamos la misma medida entre dos grupos diferentes de personas.

Existe otro tipo de diseños que requieren de la comparación entre dos medidas, pero ahora estas se han tomado en los mismos sujetos. Este es el caso de los diseños del tipo antes-después, por ejemplo, antes de haber recibido un programa educativo vs. después de haberlo hecho. Este tipo de diseños se conocen como diseños *intrasujeto*. Las mismas personas son evaluadas en diferentes momentos y se requiere la constatación de las diferencias entre estas dos etapas.

Podemos plantear diseños de investigación que, al tiempo que tienen una dimensión intersujeto, la tiene también intrasujeto. Los diseños más populares para la investigación educativa, especialmente aquellos que se usan para la evaluación de impacto, corresponden a este tipo de diseño. En ellos se diferencian varios grupos en la muestra y se examinan las variables de interés en varios momentos. El muy popular *diseño cuasiexperimental pretest/postest* (Campbell y Stanley, 1961) corresponde a este tipo.

En el diseño cuasiexperimental pretest/postest se define un grupo "experimental", en el que se aplicará un tratamiento o programa, y uno "de control", que continúa sus actividades usuales. Los resultados del programa deben ser examinados comparando las medidas de los postest entre los dos grupos. Como, con frecuencia, los grupos con los que se trabaja están previamente conformados, no se puede estar seguro de si son equivalentes respecto de sus condiciones de entrada; esta es la razón por la que aplican pruebas e instrumentos de forma previa a la implementación del programa (pretest). El diseño queda ilustrado con el esquema de la tabla 45.

Tabla 45. Esquema del diseño cuasiexperimental pretest/postest

Grupo	Pretest	Programa	Postest
Experimental	O_{1e}	X	O_{2e}
Control	O_{1c}		O_{2c}

En donde "O_{ig}" representa un episodio específico de observación y "X" representa la aplicación del programa que ocurre solo en uno de los grupos.

Existe una gran variedad de pruebas estadísticas para examinar los resultados de este experimento. Como se observa en el esquema de la tabla 45, tenemos cuatro indicadores que debemos comparar: dos en pretest y dos en postest, dos del grupo experimental y dos del grupo control. El punto es que hay aquí dos tipos de comparaciones muy diferentes.

El primer tipo de comparación se presenta entre las puntuaciones de una variable, una prueba de conocimientos, por ejemplo, entre los dos grupos, el experimental y el de control. Estamos contrastando *una variable* entre *dos subgrupos* de la muestra total y, en esa medida, es una comparación intersujeto. A las pruebas que examinan estas comparaciones se les conoce como *pruebas para dos muestras independientes.* En ellas, a la variable que contrastamos la llamamos *variable dependiente*, y a la que señala los subgrupos la llamamos *variable independiente*.

El segundo tipo de comparación se presenta entre los resultados del pretest y del postest en el mismo grupo. Este tipo de comparaciones nos resulta útil para examinar qué tanto avanzó cada grupo y si hay diferencias significativas entre el estado final y el estado inicial. En este caso, estamos comparando los resultados de dos variables diferentes (o la misma, en dos momentos diferentes), en la misma muestra y, en esa medida, es una comparación intrasujeto. Por convención, y para marcar el contraste con las anteriores, se dice que las pruebas para examinar este tipo de diferencias son *pruebas para medidas apareadas o relacionadas*. En este caso no hablamos de variables dependientes e independientes: las dos variables que se comparan están emparejadas (*paired*). La tabla 46 presenta los dos tipos de comparación.

Tabla 46. Comparaciones entre los diferentes resultados del cuasiexperimento

Tipo de prueba	Comparaciones	Significado
Pruebas para dos muestras independientes	Pretest g. experimental vs. pretest g. control	Comparamos las condiciones de entrada de los dos grupos.
Comparar los valores de *una variable* entre dos segmentos de la muestra.	Postest g. experimental vs. postest g. control	Comparamos los resultados finales de los dos grupos.
Pruebas para dos muestras apareadas / relacionadas	Postest vs. pretest en el grupo experimental	Examinamos los avances en el grupo experimental.
Comparar los valores de *dos variables* diferentes en la misma muestra.	Postest vs. pretest en el grupo control	Examinamos los avances en el grupo control.

Expondremos las pruebas para estos dos tipos de comparaciones, iniciando con las pruebas para grupos independientes. Allí examinaremos los casos en que tenemos variables métricas y continuas, ordinales y nominales.

Una vez hayamos concluido las pruebas para grupos independientes, iniciaremos la exposición de las pruebas para medidas apareadas. Como en el caso anterior, examinaremos primero las que involucran variables métricas continuas, para pasar después al caso de variables ordinales y, finalmente, al de variables con un nivel de medida nominal.

Pruebas para dos muestras independientes (una variable en dos subgrupos)

Las pruebas para contrastar los valores de una variable entre dos segmentos de la muestra se conocen como pruebas para dos muestras independientes. La selección de la prueba específica depende del nivel de medida de la variable y del cumplimiento de los supuestos de la prueba misma.

Como ya lo hemos mencionado, siempre es preferible elegir la prueba paramétrica más rigurosa; para este caso, la prueba paramétrica por excelencia para la comparación de dos muestras es la *t* de Student para grupos independientes. Esta prueba requiere que la variable dependiente sea métrica y tiene dos supuestos: la normalidad de la variable dependiente —en cada uno de los valores de la variable independiente— y la igualdad de varianzas, comúnmente llamada *homocedasticidad*.

El supuesto de normalidad puede ser el más importante para esta prueba aunque, en términos estrictos, solo se requiere que la distribución sea aproximadamente normal; esto es, simétrica y sin casos extremos. En el caso de que este supuesto no se cumpla, deben intentarse hacer las transformaciones adecuadas. Con frecuencia, cuando se hacen las transformaciones que aseguran la simetría de la variable, esta misma transformación asegura la homocedasticidad. Si esto se logra, la elección adecuada es la prueba *t* de Student para grupos independientes.

En el caso en que se logre asegurar la normalidad, pero no la homocedasticidad, la elección adecuada de la prueba es una variación de la anterior, conocida como prueba *t* de Student-Welch. En todos los programas que utilizamos se reportan los resultados de estas dos pruebas en la misma salida.

En el caso en que no sea posible asegurar la normalidad de la variable dependiente, entonces, tenemos dos posibilidades. Si la variable dependiente es métrica o si, al menos, tiene un nivel de medida ordinal, lo adecuado es una prueba no paramétrica, equivalente a la *t* de Student, conocida como la prueba U de Mahn-Witney. Si, por el contrario, la variable tiene un nivel de medida estrictamente nominal, deberemos adoptar un conjunto de pruebas basadas en el Chi-cuadrado de Pearson. El diagrama de la figura 47 describe este árbol de decisiones.

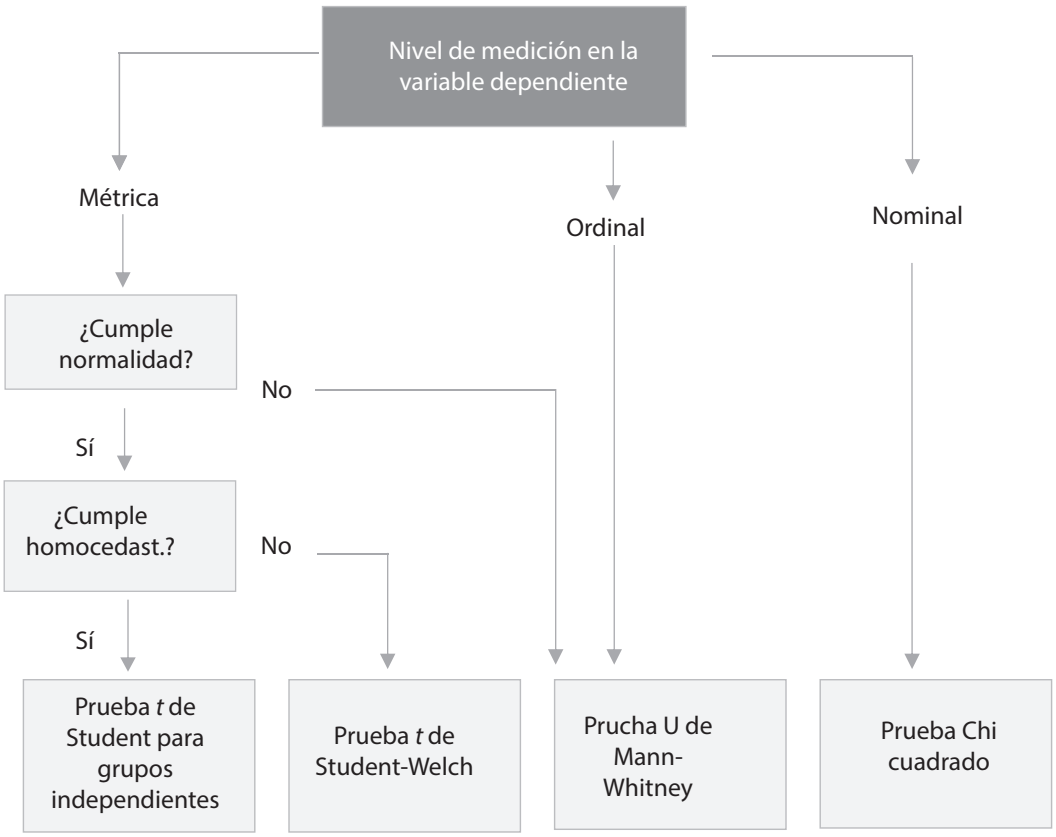


Figura 47. Pruebas para dos grupos independientes

Variable métrica: prueba t de Student para grupos independientes

Presentación general

La *prueba t de Student para grupos independientes* es una prueba paramétrica que permite examinar los niveles de significación de la diferencia entre dos medias: una obtenida en un subgrupo de la muestra y otra obtenida en otro. Requiere de una variable dependiente métrica continua (por ejemplo, el puntaje en un examen de conocimientos) y una variable independiente que contenga dos grupos disyuntos (por ejemplo, hombres y mujeres).

Esta prueba calcula el estadístico *t* y el nivel de significación asociado con este estadístico (*p*). El estadístico expresa la medida en que las diferencias entre los grupos exceden las diferencias en cada grupo. La hipótesis nula en esta prueba es que las medias de los dos grupos son iguales.

Sobre los tamaños de muestra, recomendados que se indique el uso de esta prueba en una muestra de no menos de quince individuos. En muestras grandes, el estadístico converge a una distribución normal, pero existe el peligro de que las más pequeñas variaciones en las medias muestren significancia estadística por el aumento de la potencia, razón por la cual las estimaciones de tamaño del efecto resultan muy importantes para hacer una justa valoración de las diferencias.

En términos generales, no es necesario que los dos grupos definidos por la variable independiente sean exactamente del mismo tamaño. Sin embargo, puede ser importante que estos no se encuentren demasiado desbalanceados. Por regla general, el radio entre los tamaños de los grupos debe ser menor que 1,5. Esto significa que, si un grupo tiene un tamaño n_1, el segundo grupo debe tener un tamaño n_2 tal que no salga del intervalo $[2n_1/3, 3*n_1/2]$.

Los supuestos para esta prueba son dos: 1) la normalidad de la variable dependiente y 2) la igualdad de varianzas u homocedasticidad.

En relación con el primer supuesto, se asume que la variable dependiente debe tener una distribución aproximadamente normal en cada uno de los grupos a comparar. La prueba más común para la evaluación de este supuesto es la de Shapiro-Wilk. Al respecto, se sabe que, en la práctica, aun cuando la distribución de la variable se aleje bastante de la curva normal, los resultados de la prueba *t* son bastante precisos. Por esta razón, se dice que la prueba *t* es "robusta" frente a una violación moderada del supuesto de normalidad. Como mínimo, se requiere de una distribución aproximadamente simétrica sin valores atípicos significativos. Si el presupuesto de normalidad ha sido violado de forma importante, se deben transformar los datos o, en último caso, a aplicar la prueba U de Mann-Whitney, que no requiere del cumplimiento del supuesto de normalidad.

En relación con el segundo supuesto, la homogeneidad de varianzas —también conocido como homocedasticidad— indica que las varianzas deben ser relativamente iguales en cada uno de los dos grupos. La prueba más común para la valoración de este supuesto es la prueba de Levene. En el caso de que sí se cumpla el supuesto de normalidad, pero no se cumpla el de la homocedasticidad, se puede utilizar una alternativa a la prueba *t* de Student, conocida como prueba *t* de Student-Welch o prueba de Welch.

Por último, debe señalarse algo sobre las medidas de tamaño del efecto. La medida de tamaño más común para la prueba *t* de Student o Student-Welch es la *d* de Cohen, aunque existen otras medidas relativamente conocidas que, tomando como base la anterior, corrigen algunos defectos de la *d*, tales como el delta de Glass, o la *g* de Hedges. Estas tres medidas están presentes en el JASP y en las últimas versiones del IBM-SPSS. A pesar de las ventajas de estas medidas, no resultan muy conocidas, por lo que recomendamos, por ahora, el uso de la *d* de Cohen. Para la interpretación de los resultados de la *d* de Cohen puede seguirse la tabla 47.

Tabla 47. Interpretaciones de los valores de la d de Cohen

Valor de *d*	Interpretación
$d < 0{,}2$	Efecto irrelevante
$0{,}2 < d < 0{,}5$	Efecto pequeño
$0{,}5 < d < 0{,}8$	Efecto mediano
$d > 0{,}8$	Efecto grande

Ejecutar la prueba t

Para ejecutar esta prueba en el programa JASP puede seguirse la secuencia de instrucciones recogida en el recuadro 25. Obsérvese que se han solicitado la prueba *t* de Student y la prueba de Student-Welch (Welch). Habitualmente, solo se requerirá una de ellas, pero ilustraremos las dos en el mismo ejercicio. Para correr esta prueba en el IBM-SPSS, puede seguirse la secuencia del recuadro 26. Con estas instrucciones, el programa presentará también los resultados de la prueba de Levene y los de la prueba de Student-Welch.

Recuadro 25. Solicitar una prueba t en JASP

/T-Test/Classical/ Independent Samples T-Test.
En este punto, deben seleccionarse las variables dependientes (pueden ser varias) y pasarse a la lista "Variables" y la variable independiente, pasándola a la casilla "Grouping Variable" (debe ser una variable con solo dos valores)
Test
(Seleccionar uno, dependiendo del resultado del test de igualdad de varianzas)
√ Student
√ Welch
Alt. Hypothesis
√ Group 1 ≠ Group 2
Assumption Checks
√ Normality
√ Equality of variances
Additional Statistics
√ Location parameter
√ Confidence interval [95.0%]
√ Effect Size
√ Cohen´s d
√ Descriptives
√ Descriptive plots
Confidence interval [95.0%]

Recuadro 26. Solicitar una prueba *t* en IBM-SPSS

/Analizar/Comparar medias/Prueba T para muestras independientes…
En este punto deben seleccionarse las variables dependientes (pueden ser varias) y pasarse a la lista "Variables de prueba" y la variable independiente, pasándola a la casilla "Variable de agrupación".
En el botón "Definir grupos" deben anotarse los valores a comparar
√ Estimar tamaños del efecto
Pulsar el botón "Continuar"
Pulsar el botón "Aceptar"

El ejemplo: la comparación de pruebas en grupos experimental y de control

Un profesor ha encontrado dificultades importantes en el área de Matemáticas en los alumnos del grado 10 de la institución en la cual labora. Para tratar de enfrentar esta situación, tiene la idea de diseñar cierto programa de refuerzo de Matemáticas, que debería ser implementado durante dos meses. Para probar la efectividad de este programa, después de haber recibido todas las autorizaciones correspondientes, lo ha aplicado a uno de sus grupos escolares, mientras que el otro ha continuado con sus actividades usuales, y pretende comparar los dos grupos en diferentes mediciones.

Para probar la efectividad del programa, ha elegido el diseño cuasiexperimental pretest/postest (Campbell y Stanley, 1961), en el cual se entiende la variable experimental como la exposición al programa de Matemáticas, lo cual ocurre en solo uno de los grupos: el grupo experimental. Así, este diseño define un grupo "experimental", en el que aplicará el programa, y uno "de control" que continua sus actividades usuales.

A fin de examinar las diferencias entre los grupos, se han aplicado pruebas de logro en la resolución de problemas matemáticos. Estas pruebas de logro producen información cuantitativa, métrica, a nivel de intervalo, relacionada con el número de problemas efectivamente resueltos. Utilizaremos la prueba *t* de Student para grupos independientes para el examen de las diferencias entre los grupos experimental y de control, tanto en el pretest como en el postest.

La base corresponde a datos ficticios. En total se dispone de información de 50 estudiantes, 25 conforman el grupo experimental y 25 conforman el grupo de control.

Planteamiento de las hipótesis

De acuerdo con la forma en que está planteado nuestro cuasiexperimento, esperamos que haya diferencias significativas entre las medias del postest del grupo experimental y las del grupo de control. Tal y como ya lo hemos argumentado, formularemos estas hipótesis de forma bilateral. Esto es:

> *Hipótesis nula. No hay diferencias entre las medias del postest del grupo de control y las medias del postest del grupo experimental.*
>
> *Hipótesis alternativa. Existen diferencias significativas entre las medias del postest del grupo de control y las medias del postest del grupo experimental.*

Si fuéramos más estrictos, plantearíamos hipótesis unilaterales; esto es, plantearíamos que las medias del grupo experimental deben ser estrictamente mayores que las del grupo control. Sin embargo, y como ya lo hemos justificado antes, en términos prácticos las hipótesis bilaterales son aún más exigentes, por lo que ello no es verdaderamente necesario y, en general, no se acostumbra.

Aunque las anteriores hipótesis reflejan nuestras expectativas, sabemos que hemos partido de grupos intactos, previamente conformados, y, por lo tanto, no podemos asegurar que los estados iniciales de los dos grupos sean equivalentes. Por esta razón, debemos examinar también los estados iniciales de los dos grupos para asegurar que, en efecto, los dos grupos son comparables en sus líneas de base. Así, debemos también plantear las siguientes hipótesis para el pretest.

Hipótesis nula 2: no hay diferencias entre las medias del pretest *de la prueba de Matemáticas del grupo de control y las medias del pretest del grupo experimental*

Hipótesis alternativa 2: existen diferencias significativas entre las medias del pretest *del grupo de control y las medias del pretest del grupo experimental.*

A diferencia de las hipótesis del postest, en este caso quisiéramos poder aceptar la hipótesis nula. Es decir, que los dos grupos no difieren respecto de sus estados iniciales. Esto nos simplificaría la interpretación de los resultados del postest.

Se corre la prueba

Desarrollaremos el ejemplo utilizando las salidas del programa JASP de forma simultánea para el examen de las diferencias en el pretest y en el postest.

- *Se examinan los supuestos de la prueba y se selecciona la prueba.* Como se recuerda, la prueba *t* de Student para grupos independientes requiere del cumplimiento de dos supuestos: normalidad de la variable dependiente en cada valor de la variable independiente e igualdad de varianzas. Los resultados sobre los supuestos de la prueba *t* aparecen en las tablas 48 y 49.

Tabla 48. Resultados de las pruebas de Shapiro-Wilk para examinar la normalidad de las dos variables en cada uno de los grupos, según son presentados en JASP

Test of normality (Shapiro-Wilk)			
Variable	**Grupo**	**W**	**p**
Premat	Grupo experimental	0,979	0,855
	Grupo de control	0,968	0,588
Posmat	Grupo experimental	0,967	0,571
	Grupo de control	0,932	0,097

Nota: resultados significativos sugieren una desviación de la normalidad.

La tabla 48 examina el cumplimiento del supuesto de normalidad de dos variables: premat (pretest de Matemáticas) y posmat (postest de Matemáticas). Tal y como se observa, los resultados muestran que no es posible rechazar la hipótesis nula para ninguna de las dos variables para ninguno de los dos grupos definidos por la variable independiente. Esto quiere decir que, tanto para el grupo experimental, como para el grupo de control, las distribuciones del pretest (premat) y del postest (posmat), no difieren de forma significativa de la curva normal. En conclusión, el supuesto de normalidad se cumple para los cuatro casos.

La tabla 49 muestra los resultados de la verificación del supuesto de igualdad de varianzas, u homocedasticidad. Tal y como se observa, los resultados de la prueba de Levene indican que, para ninguna de las dos variables dependientes (pretest y postest) se puede rechazar la hipótesis nula. En conclusión, el supuesto de homocedasticidad también se cumple para las dos variables.

Tabla 49. Resultados de la prueba de Levene para examinar la homocedasticidad de las dos variables

	F	df	p
Premat	0,002	1	0,963
Posmat	2,213	1	0,143

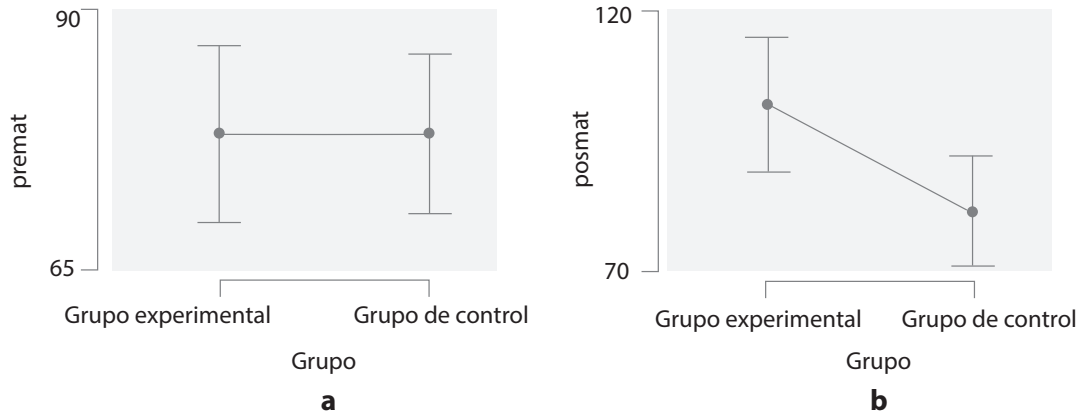
Los resultados de las pruebas nos indican que los dos supuestos requeridos para la prueba *t* de Student se verifican. Podemos seleccionar, para nuestro caso, la prueba *t* de Student para grupos independientes.

• *Se examinan los resultados descriptivos.* La tabla 50 muestra los estadísticos descriptivos, media, desviación estándar y error estándar, de cada una de las variables dependientes pretest (premat) y postest (posmat) para cada uno de los dos grupos (experimental y de control).

Tabla 50. Descriptivos del pretest (premat) y el postest (posmat) para cada uno de los dos grupos de prueba (control y experimental)

	Grupo	N	M	DE	SE
Premat	Grupo experimental	25	78,28	20,88	4,18
	Grupo de control	25	78,37	18,97	3,79
Posmat	Grupo experimental	25	102,32	31,95	6,39
	Grupo de control	25	81,32	26,13	5,22

Los resultados muestran que la media del pretest grupo experimental (con desviaciones estándar entre paréntesis) de 78,28 (20,88) es levemente más baja que la media en el grupo de control 78,37 (18,97). Por otra parte, la media del postest en el grupo experimental 102,32 (31,95) es bastante más alta que la encontrada en el grupo de control 81,32 (26,13). Los gráficos de la figura 48 indican la misma tendencia. Las barras arriba y debajo de cada media representan los errores estándar.



Nota: * las barras arriba y debajo de las medias representan el error estándar del estimador.

Figura 48. Comparación entre las medias de los dos grupos, en pretest y postest

Nota: a) diferencias en el pretest entre grupos experimental y control; **b)** diferencias en el postest entre grupos experimental y control.

Estos resultados iniciales nos harían pensar que no hay grandes diferencias en el pretest, mientras que sí se observan en el postest, a favor del grupo experimental.

Se examinan los resultados de la prueba seleccionada

La tabla 51 muestra el reporte del programa JASP sobre las dos pruebas *t* de Student corridas para examinar las diferencias entre los grupos experimental y de control.

Tabla 51. Resultados arrojados por el JASP para las dos pruebas t de Student y Student-Welch

							IC 95 % para la diferencia de medias		
	Test	**Estatistico**	**gl**	**p**	**Diferencia de medias**	**EE de la diferencia**	**Bajo**	**Alto**	***d* de Cohen**
Premat	Student	-0,02	48,00	0,986	-0,097	5,64	-11,44	11,25	-0.005
	Welch	-0,02	47,57	0,986	-0,097	5,64	-11,45	11,25	-0.005
Posmat	Student	2,54	48,00	0,014	21,005	8,25	4,41	37,60	0.720
	Welch	2,54	46,18	0,014	21,005	8,25	4,39	37,62	0.720

Como sabemos, para nuestro caso, y ya que se cumplen todos los supuestos, la prueba adecuada es la prueba *t* de Student para grupos independientes. Los resultados aparecen en los primeros renglones después de la variable.

Iniciando con los resultados del pretest (premat), la prueba indica que no hay diferencias significativas entre las medias de los dos grupos. Esto se confirma cuando constatamos que los límites del intervalo de confianza para la diferencia de medias al 95 % contienen el valor 0 dentro del intervalo, por lo que no se permite el rechazo de la hipótesis nula. Correspondientemente con ello, el valor del tamaño del efecto (*d* de Cohen) es casi nulo.

Al contrario del pretest, el postest (posmat) muestra diferencias significativas en las medias de los dos grupos (p=,014<,05). Correspondientemente con ello, el intervalo de confianza para la diferencia de las medias al 95 % no incluye en valor 0. Por estas razones, es posible rechazar la hipótesis nula y adoptar la hipótesis alternativa. El tamaño del efecto, indicado por la *d* de Cohen, por su lado, muestra ser bastante grande.

Puede ser importante notar que el signo del valor *t* es negativo para el pretest y positivo para el postest. Esto indica únicamente el sentido de la diferencia. En la medida en que habíamos asumido que el grupo experimental tenía el valor "2" y el de control el valor "1" y se calculó la diferencia entre las medias en términos de la media del grupo experimental menos las del grupo de control. En el pretest esto era levemente negativo y en el postest marcadamente positivo. De haberlos definido al contrario, los signos de los valores *t*, los límites de los intervalos de confianza y la *d* de Cohen se hubieran invertido. Esto no tiene mayores implicaciones para la interpretación de los resultados.

Si no se hubiera cumplido el supuesto de igualdad de varianzas, hubiéramos debido tomar el resultado de la prueba *t* de Student-Welch, cuyos resultados aparecen como "Welch" en el segundo renglón después de la variable dependiente. Puede ser interesante observar que, en comparación con la *t* de Student, la prueba *t* de Student-Welch solo difiere levemente en algunos valores: los

grados de libertad (gl) y el error estándar, lo que también modifica levemente los intervalos de confianza. Aunque habitualmente las conclusiones son idénticas, al reportar los resultados es importante tener en cuenta la prueba que resulte apropiada para el caso.

Se expresan los resultados

Para la expresión de los resultados de las pruebas *t* de Student, existe un formato específico que permite exponer, con claridad y precisión, los principales resultados en el texto. Para el caso de la prueba t sobre muestras independientes, este formato es el que se presenta en la figura 49.

t(<gl>*) = <valor *t*> *p* = <valor *p*> *d* = <valor *d*>

***Nota:** gl: grados de libertad;
Los valores entre los signos <> son los arrojados por la prueba.

Figura 49. Formato para expresar los resultados de las pruebas t de Student sobre grupos independientes

Siguiendo este formato, los resultados de las dos pruebas pueden ser expresados en texto de la siguiente forma:

> *El examen de los supuestos de normalidad y homocedasticidad mostró la pertinencia de utilizar la prueba t de Student para grupos independientes: tanto las pruebas de Shapiro-Wilk y como la de Levene muestran que la distribución no difiere de forma significativa de la curva normal y las varianzas puede ser consideradas iguales. Los resultados de la aplicación de esta prueba muestran que la diferencia media entre los grupos experimental y de control en el pretest de Matemáticas no fue significativa t(48) = 0,10 p = -,986 (NS) d = 0,01. Por el contrario, los resultados de la misma prueba muestran la diferencia media entre estos grupos en el postest fue significativa a nivel de ,05 con un tamaño del efecto mediano t(48)=2,54 p =,014 d = 0,72.*

Otra posibilidad de expresar los resultados es hacerlo en una tabla (tabla 52), que permitiría, de igual forma, incluir otros datos, tales como la diferencia media entre los grupos y el intervalo de confianza para la diferencia.

Tabla 52. Resultados de una prueba t de Student expresados en una tabla

Prueba	Grupo	*M*	*DE*	*t*	*gl*	*p*	*d* de Cohen
Pretest	Experimental	78,28	20,88	-0,10	48	,986	-0,01
	Control	78,37	18,97				
Postest	Experimental	102,32	31,95	21,00	48	,014*	0,72
	Control	81,32	26;13				

*: $p<,05$.

Para terminar, debe anotarse que todo lo dicho para la prueba *t* de Student es igualmente cierto para la prueba *t* de Student-Welch, o prueba de Welch, utilizada cuando el supuesto de homocedasticidad no puede ser verificado. Para este caso, las salidas del programa mostrarán una leve diferencia en los grados de libertad y en los límites de los intervalos de confianza. Los resultados se expresan de la misma forma que en la prueba *t* de Student usual.

Variable ordinal: prueba U de Mann-Whitney

Presentación general

Cuando hay graves violaciones al supuesto de normalidad, cuando las diferencias entre los tamaños de los dos grupos son muy extremas o cuando la variable dependiente tiene un nivel de medida ordinal, la prueba adecuada para el examen de diferencias entre dos grupos es el equivalente no paramétrico de la prueba *t*: la *prueba U de Mann-Whitney*.

La prueba U de Mann-Whitney es una prueba de rangos. Eso significa que trabaja con posiciones, más que con números; perfecto para las variables ordinales. En la medida en que trabaja con variables con niveles de medida ordinal, evita la comparación de medias, privilegiando la comparación de rangos promedio (o las medianas de los valores no convertidos a rango).

La prueba U de Mann-Whitney requiere de una variable dependiente, cuyo nivel de medida debe ser, al menos, ordinal, y una variable independiente que defina dos grupos disyuntos. Fuera de estas dos condiciones básicas, no requiere del cumplimiento de ningún supuesto. También produce un estadístico, *U*, y un nivel de significación asociado a este estadístico (*p*) para la diferencia entre los rangos promedio de los dos grupos.

Dependiendo del *software* estadístico que se utilice, las salidas pueden cambiar de forma notable. En todos los casos se obtiene un estadístico (*U*, o a veces *W* de Wilcoxon), siempre positivo y el nivel de significación asociado. En el caso del SPSS, se produce una tabla en la que se presentan los rangos promedio de la variable dependiente en cada uno de los grupos que se comparan, lo que permite examinar el sentido de las diferencias. En otros programas, como el JASP, la prueba sigue mostrándose con tablas y gráficos de medias y errores estándar, lo que formalmente no resulta muy adecuado, pero usualmente resulta suficiente para determinar la dirección de los cambios.

Otra de las variaciones más importantes entre los programas es el uso de estimadores de la diferencia entre los dos grupos. Para el caso de JASP, se presenta un cierto estimador específico conocido como el *estimador de Hodges-Lehmann*, el cual es robusto y no paramétrico basado en la prueba de los rangos con signo de Wilcoxon, es utilizado para estimar diferencias entre dos poblaciones. El SPSS no muestra un estimador de la diferencia entre los grupos.

Al respecto de las medidas de tamaño del efecto también se dan diferencias importantes entre los programas. El IBM-SPSS no presenta ninguna medida de tamaño del efecto, por lo que las personas que lo utilicen deberán calcular esta medida en calculadoras de tamaño del efecto o aplicaciones especializadas, como la página www.psychometrica.de. Esta página tiene una sección específica para el cálculo del tamaño del efecto en pruebas no paramétricas.

Por su parte, en el programa JASP se ofrece, como medida para el cálculo del tamaño del efecto, la *correlación rango-biserial* (r_{rankb}). Esta es la medida de asociación más adecuada para los casos en que tenemos una variable ordinal y una variable dicotómica. Ya que representa una medida de asociación, puede ser entendida también como una medida de tamaño del efecto. Para su interpretación deben utilizarse las mismas reglas de Cohen (1988) para la interpretación de la correlación de Pearson (*r*) como medida de tamaño del efecto, expuestas en la tabla 53.

Tabla 53. Interpretación de los valores de r o r_b como medidas de tamaño del efecto

r_b	Interpretación, según Cohen (1988)
r_b <0,1	Sin efecto
0,1 < r_b <0,3	Efecto pequeño
0,3 < r_b <0,5	Efecto medio
0,5 < r_b	Efecto grande

Fuente: Cohen (1988).

Cómo ejecutar la prueba U de Mann-Whitney

Para ejecutar la prueba U de Mann-Whitney en el programa JASP, puede seguirse el camino reflejado en el recuadro 27. En el IBM-SPSS, es posible encontrar esa prueba en /Analizar/Pruebas no paramétricas (recuadro 28). El IBM-SPSS no aportará información sobre medidas de tamaño del efecto.

Recuadro 27. Cómo ejecutar la prueba U de Mann-Whitney en JASP

/T-Test/Classical/ Independent Samples T-Test.
En este punto, deben seleccionarse las variables dependientes (pueden ser muchas) y pasarse a la lista "Variables" y la variable independiente, pasándola a la casilla "Grouping Variable" (debe ser una variable con solo dos valores).

Test
(Seleccionar uno, dependiendo del resultado del test de igualdad de varianzas).
√ Mann-Whitney
Alt. Hypothesis
√ Group 1 ≠ Group 2
Additional Statistics
√ Location parameter
√ Confidence interval [95,0 %]
√ Effect Size
√ Cohen´s d
√ Descriptives
√ Descriptive plots
Confidence interval [95,0 %]

Recuadro 28. Cómo ejecutar la prueba U de Mann-Whitney en IBM-SPSS

/Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/Dos muestras independientes…
En este punto deben pasarse una, o varias, variables dependientes a "Lista de variables de prueba" y la variable independiente a "Variable de agrupación"
En el botón "Definir grupos"
Definir los dos valores a comparar
Pulsar el botón "Continuar"
√ U de Mann-Whitney
Pulsar el botón "Aceptar"

El ejemplo: diferencias en las actitudes en el pretest y en el postest

En el ejemplo anterior, en el que el investigador desea determinar el efecto de un programa de Matemáticas a través de un diseño cuasiexperimental pretest/postest, se quiso complementar la observación de las pruebas haciendo una medición de las actitudes presentes hacia las materias.

Para hacerlo, el investigador ha aplicado cuestionarios a toda la muestra para indagar acerca de *sus actitudes hacia la materia* de forma previa y posterior al programa de Matemáticas, tanto en el grupo experimental como en el de control. En este caso, el estudiante califica su propia actitud frente a la materia en una *escala ordinal* de cuatro puntos, en donde "1" representa una actitud negativa, "2" una actitud neutra, "3" una actitud positiva y "4" una actitud muy positiva hacia la materia.

De nuevo tenemos, frente a la actitud, la misma situación que la presentada frente a la prueba. Se espera que los estudiantes del grupo experimental muestren en el postest actitudes más positivas hacia la materia que los estudiantes del grupo control. Se supone que no debe haber diferencias previas en las actitudes entre los dos grupos, pero, en la medida en que ello no puede asegurarse, se examinaran las diferencias eventualmente presentes en el pretest de actitudes.

Planteamiento de las hipótesis

Las hipótesis en una prueba de rangos no son exactamente iguales a las hipótesis en una prueba paramétrica. La diferencia básica es que en una prueba paramétrica, como la *t* de Student, se comparan las medias de dos grupos, y por tanto la hipótesis nula plantea que los dos grupos tienen la misma media. En una prueba de rango, no podemos obtener la media, sino que su equivalente sería el rango medio, o la mediana. En esa medida, la hipótesis nula de la prueba debe ser formulada en términos de la igualdad de las medianas.

Así, las hipótesis sobre las diferencias en el postest quedarían formuladas como sigue:

> *Hipótesis nula 1. No existen diferencias entre las medianas del postest de actitudes hacia el programa de Matemáticas entre los estudiantes del grupo experimental y los del grupo de control.*

> *Hipótesis alternativa 1. Existen diferencias entre hombres y mujeres en las medianas del postest de actitudes hacia el programa de Matemáticas entre los estudiantes del grupo y los del grupo de control.*

Al probar estas hipótesis estaremos examinando si, después de la aplicación del programa, habrá diferencias entre los dos grupos.

De nuevo, debemos examinar las diferencias en las actitudes previas a la aplicación del programa entre los dos grupos. Estas hipótesis quedarían formuladas de la siguiente forma.

> *Hipótesis nula 2. No existen diferencias entre las medianas del pretest de actitudes hacia la matemática entre los estudiantes del grupo experimental y los del grupo de control.*
>
> *Hipótesis alternativa 2. Existen diferencias entre hombres y mujeres en las medianas del pretest de actitudes hacia matemáticas entre los estudiantes del grupo experimental y los del grupo de control.*

Al examinar estas hipótesis estaremos examinando si, de forma previa a la aplicación del programa, había diferencias en las actitudes presentes entre los dos grupos que pudieran contribuir a explicar los resultados posteriores.

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* La prueba U de Mann-Whitney solo requiere que la variable dependiente —la medida de la actitud— tenga un nivel de medida ordinal y que la variable independiente defina dos grupos. Estos dos presupuestos se cumplen. La variable dependiente tiene un nivel de medida ordinal, de cuatro puntos (actitud negativa, neutra, positiva y muy positiva) y el diseño define dos grupos independientes (experimental y control). Podemos aplicar la prueba U de Mann-Whitney.
- *Se examinan los resultados descriptivos.* Para este tipo de situaciones, en que tenemos variables ordinales en dos muestras independientes, lo indicado para hacerse una idea clara de las similitudes y diferencias entre los dos grupos es hacer tablas de cruce (*crosstabs*) entre las variables de actitudes y grupo, por separado para cada una de las pruebas. En la tabla 54 se muestran los resultados de estos cruces.

Tabla 54. Cruce entre actitudes y grupo, para cada prueba

Prueba	Grupo	Actitud hacia la matemática				Total
		Negativa	Neutra	Positiva	MPositiva	
Pretest	Control	5	7	7	6	25
	Experimental	5	12	7	1	25
	Total	10	19	14	7	50
Postest	Control	3	7	11	4	25
	Experimental	0	5	11	9	25
	Total	3	12	22	13	50

Los resultados de estas tablas pueden ser fácilmente representados en una gráfica de barras apiladas en el que se han igualado los grupos al 100 %. El resultado para la comparación entre los dos grupos en pretest y postest se presenta en la figura 50.

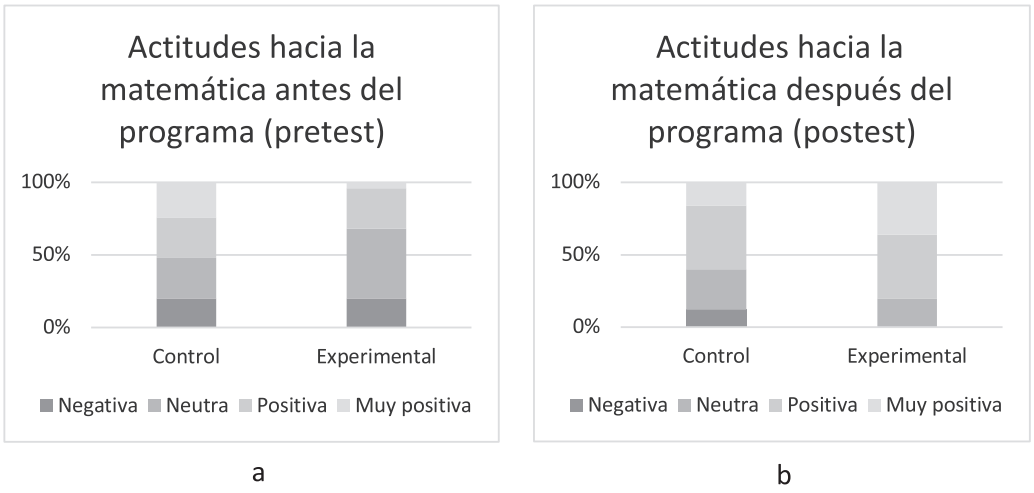


Figura 50. Gráficas del cruce entre las actitudes frente a las matemáticas por grupo con columna 100 % apilada.
Notas: a) pretest; b) postest.
Fuente: elaboradas en Excel sobre las tablas de cruce.

Para la interpretación, basta con inspeccionar visualmente las áreas en las que se presentan cambios entre los dos grupos. Procediendo de esta forma, observamos que, en el caso de pretest (figura 50a), el grupo experimental muestra una proporción más grande de actitudes neutras y una proporción más baja de actitudes muy positivas. Aunque no sabemos qué tan significativa es la diferencia, podría sugerir que las actitudes de entrada en el grupo experimental podrían ser menos favorables que las del grupo de control.

El examen de la figura 50b, donde se comparan los valores del postest, muestra que el grupo experimental manifiesta menor proporción de actitudes negativas y neutras y mayor proporción de actitudes muy positivas. De nuevo, sin que aún sepamos qué tan significativas sean estas diferencias, parece indicarse que el grupo experimental expresa, en el postest, actitudes más positivas hacia la materia que el grupo de control.

Aunque lo anterior es, formalmente, lo más adecuado, es un proceso relativamente largo y dispendioso que, en general, no resulta estrictamente necesario. En la figura 51 se muestran las gráficas de la comparación de medias, con errores estándar por encima y por debajo, de las actitudes por grupos de género, tal y como es mostrado por el programa JASP. Como se observa, el resultado es igual al obtenido con los procedimientos anteriores: mientras que, en el pretest el grupo de control muestra mejores actitudes hacia el programa de Matemáticas que el experimental, en el postest se observa lo contrario, y con mayor intensidad: el grupo experimental muestra mejores actitudes que el de control.

Lamentablemente, este modo de proceder no es formalmente adecuado, en la medida en que estamos comparando medias de variables ordinales y, como se recordará, no estamos estrictamente autorizados para utilizar la media como medida de tendencia central en variables ordinales; más adecuada sería la mediana. Sin embargo, la mediana no siempre se diferencia lo suficiente entre los grupos. Por esta razón, es más frecuente el uso de la media para examinar rápidamente las comparaciones de variables ordinales entre dos grupos.

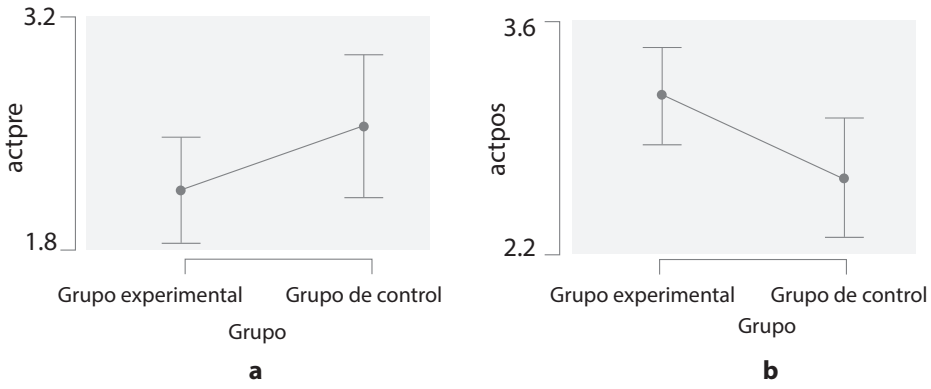


Figura 51. Medias y errores estándar del puntaje de actitud para cada grupo.
Nota: a) en el pretest; b) en el postest.

Existe una última posibilidad de examinar de forma gráfica las diferencias atendiendo al nivel de medida, ordinal, de la variable dependiente: una gráfica de cajas y bigotes (*box plots*). Estas gráficas son adecuadas para variables ordinales. La figura 52 muestra este tipo de gráficas para el pretest y postest, tal y como son generadas por el programa JASP.[5]

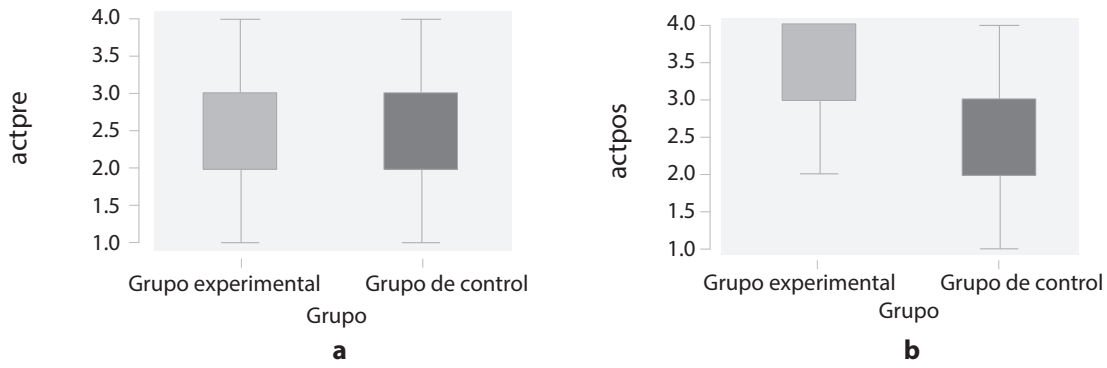


Figura 52. Gráficas de cajas y bigotes del puntaje de actitud para cada grupo.
Nota: a) en el pretest; b) en el postest.

Como se observa, estas gráficas no resultan tan precisas como las barras apiladas o las gráficas de medias. De acuerdo con ellas, en el pretest no parece haber diferencias entre los grupos (figura 52a), mientras que en el postest (figura 52b) el grupo experimental muestra mejores actitudes que el grupo de control.

Mucho más adecuado a la naturaleza de la prueba que vamos a examinar es analizar el *rango promedio* entre los dos grupos, a fin de estimar la dirección de las influencias. Sin embargo, no todos los programas hacen explícito el rango promedio, incluso cuando corren pruebas de rangos. En el JASP esta información no aparece disponible. La tabla 55 muestra los resultados del SPSS cuando se corren pruebas U de Mann-Whitney, con estos datos.

5 Para generar este tipo de gráficas en JASP, es necesario cambiar el nivel de medida explícito de la variable, de ordinal a métrico. Una condición que no debería ser necesaria pero es requisito para poder hacerlo.

Tabla 55. Salida del spss cuando se corren prueba U de Mann Whitney de las actitudes entre los grupos experimental y de control en el pretest y postest

Rangos				
	Grupo	**N**	**Rango promedio**	**Suma de rangos**
Pretest de actitud	**Grupo de control**	25	28,20	705,00
	Grupo experimental	25	22,80	570,00
	Total	50		
Postest de actitud	**Grupo de control**	25	21,60	540,00
	Grupo experimental	25	29,40	735,00
	Total	50		

De acuerdo a lo que se observa en la tabla, los rangos promedio de la segunda columna muestran la misma tendencia que habíamos visto antes para las diferencias entre los grupos. Mientras en el pretest, el rango promedio del grupo de control es más alto que el del grupo experimental, en el postest ocurre lo contrario, y el rango promedio de las actitudes en el grupo experimental es más alto que el del grupo de control. La medida de la significación de estas diferencias es lo que deberemos examinar ahora.

Se examinan los resultados de la prueba

Para obtener los resultados de la prueba U de Mann Whitney en el programa JASP, debe seguirse este camino: T-Test > Classical > Independient Samples T-Test. En este punto deben solicitarse, estrictamente, pruebas U de Mann Whitney. Existen otras opciones, dentro de las cuales hemos seleccionado el valor de parámetro de localización, con intervalo de confianza al 95 % y una estimación del tamaño del efecto. La tabla 56 muestra los resultados, según los presenta el programa JASP.

Tabla 56. Resultados del programa JASP con las pruebas U de Mann Whitney las diferencias entre las actitudes entre los dos grupos de forma separada para el pretest y postest

				95 % CI para estimado de Hodges-Lehmann		
	W	**p**	**Estimado de Hodges-Lehmann**	**Bajo**	**Alto**	**Correlación Rango-Biserial**
Actpre	245,0	0,174	-4,272e -5	-1,000	7,857e -5	-0,216
Actpos	410,0	0,045	1,000	-2,271e -5	1,000	0,312

Note. For the Mann-Whitney test, effect size is given by the rank biserial correlation.

Note. Mann-Whitney U test.

Primero lo básico: el examen de la significación estadística. Analizando primero el pretest (Actpre), los resultados muestran que no hay diferencias estadísticamente significativas, a nivel de 0,05, en

las actitudes de los dos grupos en el pretest. En efecto, los niveles de *p*, aunque bajos, no alcanzan a ser más bajos que nuestro punto de corte elegido por defecto (*p*=,174 > ,050).

El examen de las diferencias en el postest, por otro lado, manifiesta que sí hay diferencias significativas entre los dos grupos, en la medida en que estas son inferiores al punto de corte previamente elegido (*p*=,045 < ,050). Esto nos autoriza a rechazar la hipótesis nula y a adoptar la alternativa: las actitudes del grupo experimental después del programa muestran ser más positivas que las del grupo de control.

Los resultados contenidos de la tabla 56 introducen dos nuevos indicadores que debemos explicar. Primero, el parámetro de localización de la diferencia entre los dos grupos, que ahora corresponde a la "estimación de Hodges-Lehmann". Este parámetro pertenece a la diferencia mediana entre los dos grupos. El intervalo de confianza para este estimador se interpreta como es usual. En el pretest el estimador queda por dentro del intervalo, lo que muestra la ausencia de diferencias, y en el postest queda por fuera, en realidad en el borde, lo que manifiesta que la diferencia es significativa.

Segundo, la estimación del tamaño del efecto, que ahora corresponde a la "correlación de rango biserial" (*rank biserial correlation*), r_{rankb}, o r_b, puede ser considerada como una medida de tamaño del efecto apropiada para esta prueba. Para la interpretación de la correlación rango biserial, como medida de tamaño del efecto, puede recurrirse a la forma en la que se interpreta la correlación *r* de Pearson. En ese sentido, las correlaciones r_b menores a ,3, presentes en la diferencia entre los grupos sobre las actitudes en el pretest deben ser interpretadas como un pequeño tamaño de efecto; mientras el r_b=,312, en la diferencia entre los grupos en el postest, indica un tamaño del efecto que puede ser considerado intermedio.

El reporte de esta prueba en el IBM-SPSS es bastante diferente. La tabla 57 muestra la forma en que este programa presenta los resultados. Se exhibe el valor U, el W de la prueba de Wilcoxon y la tipificación de los dos valores (Z). Como se observa, aunque los valores de los estadísticos difieren en las salidas de los dos programas, los niveles de significación son muy similares.[6]

Tabla 57. Reporte de la prueba U de Mann Whitney en el ibm-spss

Estadísticos de prueba[a]		
	Pretest de actitud	**Postest de actitud**
U de Mann-Whitney	245,000	215,000
W de Wilcoxon	570,000	540,000
Z	-1,371	-2,013
Sig. asintótica (bilateral)	,170	,044

a. Variable de agrupación: grupo.

6 Aparentemente, la razón de que los valores de las pruebas difieran según el paquete estadístico utilizado está relacionada con decisiones específicas en que los procedimientos difieren, sin que ello afecte en mayor medida los niveles de significancia. Específicamente hemos encontrado que, mientras en algunos paquetes (SPSS, por ejemplo) se toma como el valor de la prueba la suma de rangos sin empates negativos, en otros se toman la suma de rangos sin empates positivos. El cálculo del nivel de significación se hace, en los dos casos, con base en la aproximación a la distribución normal estándar. Agradezco al profesor Carlos Lanziano el señalamiento de este tipo de diferencias entre los paquetes.

Puede ser interesante anotar que los valores de las significaciones encontradas en las dos pruebas, así como los de tamaño del efecto, no muestran ser extremadamente concluyentes. En el pretest, aunque la diferencia no es significativa, el valor no dista mucho de serlo. En el postest, aunque la diferencia es significativa, el valor de la significación se encuentra muy cerca del límite definido (0,050). Esto debe alertarnos para interpretar con cuidado los resultados, ya que, como sabemos, los límites de significación son convenciones históricas más que valores absolutos.

Desde otro punto de vista, la constatación de que se parte de actitudes levemente más positivas en un grupo y se termina con actitudes marcadamente más positivas en el otro podría considerarse como un apoyo adicional a la idea de que el programa de Matemáticas tiene un efecto en la modificación de las actitudes hacia la materia en los estudiantes que lo toman.

Se expresan los resultados

No hay un consenso muy claro acerca de cómo deben ser presentados los resultados de la prueba U de Mann Whitney. Algunos anotan que deben anexarse todos los datos de los valores U, W y Z, además de los niveles de significación. En realidad, todos estos valores resultan equivalentes. Por esta razón, y en aras de no sobre cargar el texto con estadísticas que resultan redundantes, sugerimos incluir los valores U, los niveles de significación y alguna medida de tamaño del efecto, preferiblemente la correlación rango-biserial, de la forma

U=<valor U> p = <valor p> r_b= <valor de la correlación rango biserial>

siguiendo las recomendaciones sobre itálicas y cantidad de decimales en cada medida.

Los resultados obtenidos pueden ser expresados, en el texto, de la siguiente forma:

> *Los resultados muestran que, para el caso del pretest, que aunque la medida de las actitudes frente a la materia es levemente más favorable en el grupo de control comparado con el grupo experimental, esa diferencia no es significativa en una prueba U de Mann-Whitney a nivel de ,05, U = 245,00 p =, 174 (ns)* r_b*=-,216. En el caso del postest, por el contrario, el grupo experimental manifiesta actitudes hacia la materia más favorables que el grupo de control, la diferencia es significativa y el tamaño del efecto puede ser considerado intermedio U = 410,00 p= ,045* r_b*=,312.*

Como se observa, no hemos incluido el parámetro de localización de la diferencia mediana entre los grupos (estimador de Hodges-Lehmann) ni su intervalo de confianza asociado. La razón de no hacerlo es el relativo desconocimiento que se tiene hoy sobre este indicador. Si en algún momento su uso se generaliza, valdrá la pena incluirlo en los textos científicos. De otra forma, exigiría dar demasiadas explicaciones que no parecen ser estrictamente requeridas por las publicaciones en el momento presente.

Otra forma de presentar los resultados podría ser mediante una tabla, y muy especialmente si se deben presentar junto con los resultados de otras muchas pruebas. Un ejemplo es la tabla 58.

Tabla 58. Resultados de las pruebas U de Mann-Whitney de diferencias entre el grupo experimental y el de control en el presente y postes

Variable	*U*	*p*	r_b
Pretest	245,00	,171	-,216
Postest	410,00	,045*	,312

Nota: *: ,05<*p*

De cualquier forma, para expresar los resultados debe incluirse, siempre, uno o varios de los estadísticos de la prueba (*U, W, Z*), el nivel de significación encontrado (*p*) y alguna de las diferentes medidas de tamaño del efecto (*d,* r_b,…).

Variable nominal: prueba Chi-cuadrado (χ^2) de Pearson

Presentación general

Continuamos la comparación entre dos grupos, ahora con la situación en que la variable dependiente es de naturaleza nominal. Para esta situación contamos con un conjunto de pruebas basadas en el estadístico Chi-cuadrado (χ^2). En este tipo de pruebas se parte del examen de la tabla de cruce entre las dos variables y se compara una distribución *observada* de los datos con la distribución *esperada* para los mismos, si las variables fueran independientes entre sí.

La hipótesis nula para una prueba χ^2 *de Pearson* es la de la independencia de las dos variables. Si las dos variables son independientes, esto significa que no presentan asociación y que, por lo tanto, los valores de una no dependen de los valores de la otra. La negación de esta hipótesis consiste en afirmar que los valores de una de ellas dependen de los valores de la otra.

Para este tipo de prueba es importante distinguir dos tipos de estadísticos por su función: los estadísticos de prueba de hipótesis de independencia y las medidas de asociación que permiten estimar el tamaño del efecto.

Al respecto de los estadísticos para la prueba de la hipótesis de independencia, la prueba χ^2 de Pearson produce un estadístico (χ^2), una estimación del grado de libertad (gl) y un valor de probabilidad asociado con ese estadístico (*p*). El valor de χ^2 no tiene límite superior; varía entre 0 e infinito y depende, en parte, del tamaño de la muestra. A mayor tamaño de la muestra, mayor el valor del χ^2.

Existen unas restricciones importantes para el uso de la prueba χ^2 relacionadas con el tamaño de la muestra y la distribución de la misma en cada una de las casillas de la tabla de cruce. Específicamente, debe anotarse que el uso de esta prueba no se recomienda en muestras pequeñas (n<30), o en situaciones en que exista una celda dentro del cruce con una frecuencia inferior a cinco casos o, peor aún, una celda vacía.

Para el caso de muestras pequeñas, se recomienda una modificación a la prueba χ^2 que se conoce como *corrección por continuidad*, que se utiliza para prevenir una sobreestimación de la significación; el problema con esta corrección es que resulta muy conservadora y ahora se cae en el efecto

contrario. Otra variación para muestras pequeñas es la conocida como *prueba exacta de Fisher*, pero esta se diseñó solo para tablas de 2x2, si bien algunos opinan que puede ser utilizada en tablas de mayor dimensión.

En general para este caso, el *software* estadístico es bastante oscuro. La impresión para el neófito es de una enorme dispersión de pruebas con muy poca información acerca de las circunstancias en que estas pueden ser efectivamente utilizadas.

A pesar de toda esta confusión, todo parece indicar que es posible recomendar una prueba, dentro de la familia de pruebas χ^2, que puede servirnos en todas las situaciones: la *razón de verosimilitud (likelihood radio).* Esta prueba es una alternativa a la χ^2 basada en la teoría de máxima verosimilitud. Aunque se recomienda para muestras de tamaño pequeño (n<30), en muestras grandes produce el mismo resultado que el χ^2 de Pearson. Aparentemente esta prueba permite su uso en todas las circunstancias, dimensiones y tamaños de muestra posibles y se encuentra disponible en todos los paquetes estadísticos de uso común, y en particular los que hemos ilustrado en este libro (JASP y IBM-SPSS).

Al respecto de las medidas para estimar el tamaño del efecto, la situación también resulta un poco confusa. Dependiendo del paquete estadístico, el tamaño de la muestra y el número de dimensiones del cruce de las variables se mencionan, como medidas de tamaño del efecto, el coeficiente Phi (ϕ), la *V* de Cramer, el coeficiente de contingencia y el coeficiente de incertidumbre. Las restricciones para el uso de cada uno de estos están siempre presentes y no siempre son claras. El coeficiente ϕ, por ejemplo, solo puede ser usado en tablas 2x2; el coeficiente de contingencia, por su parte, sólo se recomienda en tablas de dimensión 5x5 o más.

La medida de tamaño del efecto más popular, más usada y con menor número de restricciones es una de asociación basada en Chi-cuadrado conocida como la *V de Cramer*, simbolizada, en ocasiones, mediante el símbolo ϕ_C (Phi de Cramer) o V_C. El valor de la *V* de Cramer se encuentra entre 0 (no hay relación) y 1 (relación perfecta). La interpretación del valor específico encontrado, sin embargo, cambiará dependiendo del número de grados de libertad de la tabla de cruce. La tabla 59 presenta los valores de tamaño del efecto dependiendo de los grados de libertad de la tabla de cruce.

Tabla 59. Interpretación de los valores de Phi y V dependiendo de los grados de libertad

Tamaño del efecto	gl* (df)	Pequeño	Mediano	Grande
ϕ y *V* de Cramer	1	0,10	0,30	0,50
V de Cramer	2	0,07	0,21	0,35
	3	0,06	0,17	0,29
	4	0,05	0,15	0,25
	5	0,04	0,13	0,22

* Los grados de libertad (gl) dependen de la dimensión de la tabla. En una tabla de dimensión r*c, los grados de libertad serán gl= (r-1)*(c-1)

Fuente: adaptado de Kim (2017) y Goss-Sampson (2019).

Para el caso específico que nos ocupa en este apartado, tenemos una variable "dependiente", de naturaleza nominal, y una variable "independiente" que define los dos grupos que deseamos comparar. El entrecomillado en las variables subraya el hecho de que la prueba es indiferente a cuál de las variables se encuentra en las filas o en las columnas: el resultado es idéntico. En todo caso, la variable que hemos rotulado como "independiente" definirá dos grupos, por lo que los grados de libertad de la tabla serán iguales al número de filas menos 1.

Cómo ejecutar la prueba Chi-cuadrado (χ^2) de Pearson

Para el cálculo de estas pruebas, puede procederse, en JASP, de la forma presentada en el recuadro 29; en el IBM-SPSS, como en el recuadro 30.

Recuadro 29. Cómo ejecutar la prueba Chi-cuadrado en JASP

/Frecuency/Contingency Tables

En este punto, deben seleccionarse la variable dependiente (y pasarse a la lista "Filas" y la variable independiente, pasándola a la casilla "Columnas" (en esta sección debería ser una variable con solo dos valores pero admitirá cualquier variable categórica).

Statistics (debe seleccionarse la prueba adecuada al tamaño de muestra).

√ χ^2

√ χ^2 continuity correction

√ Likelihood radio

Nominal

√ Phi and Cramer's V

Recuadro 30. Cómo ejecutar la prueba Chi-cuadrado en IBM-SPSS

/Analizar/Estadísticos descriptivos/Tablas cruzadas…

En este punto, deben seleccionarse la variable dependiente (y pasarse a la lista "Filas" y la variable independiente, pasándola a la casilla "Columnas" (para esta sección debería ser una variable con solo dos valores, pero admitirá cualquier variable categórica).

En el botón "Estadísticos"

√ Chi-cuadrado

Nominal

Phi y V de Cramer

pulsar el botón "Continuar"

Pulsar el botón "Aceptar"

El ejemplo: ¿hay diferencias en los tipos de familia de estudiantes que desertan y los que no?

Se ha notado que los estudiantes que muestran ciertos niveles de deserción parcial tienen mayores probabilidades, en el futuro, de desertar completamente del sistema educativo. Por esta razón, unos investigadores han planteado la necesidad de hacer un perfil con las características distintivas de los estudiantes que han tenido episodios de deserción parcial a lo largo de su historial académico.

Una de las características que podría estar asociada con este tipo de deserción es la tipología de familia. En particular, los investigadores opinan que en familias intactas nucleares y completas, los estudiantes tienen menores posibilidades de desertar del sistema educativo.

Para esta indagación, se categoriza el tipo de familia, como una variable nominal con cuatro categorías, a saber:

- Familia sin padre ni madre.
- Familia monoparental paterna.
- Familia monoparental materna.
- Familia nuclear (con padre y padre).

Así, se pretende examinar las diferencias familiares entre los estudiantes que han desertado del sistema educativo, al menos una vez, por algún periodo de tiempo, y aquellos que no lo han hecho. Tenemos acá la indagación de diferencias entre estudiantes que han desertado y los que no lo han hecho, respecto de los valores observados en una variable nominal politómica: la tipología de familia.

Para este ejemplo contamos con datos efectivamente recogidos en un proceso de investigación llevado a cabo hace algunos años. En total, contamos con información proveniente de 1499 estudiantes de grado décimo de colegios públicos de Bogotá.

Planteamiento de las hipótesis

La prueba χ^2 de Pearson es una prueba de independencia. En esa medida, la hipótesis nula indicará la independencia de las dos variables, mientras que la hipótesis alternativa constatará su dependencia, lo que es lo mismo que la diferencia significativa entre los valores de una variable dependiendo de los valores de la otra.

En el ejemplo que desarrollaremos, los investigadores opinan que la deserción parcial del sistema educativo tiene relación con la estructura de la familia y, en particular, con el hecho de que la familia del estudiante muestre una estructura nuclear completa, esto es, con padre y madre presentes. Así las cosas, las hipótesis pueden ser formuladas como sigue.

> *Hipótesis nula (H_0). La tipología de familia en los estudiantes desertores es igual a la tipología de familia de los estudiantes que no han desertado del sistema educativo.*
>
> *Hipótesis alternativa (H_1). Existen diferencias en las tipologías de familia de los estudiantes desertores y las de los estudiantes no desertores.*

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* Las pruebas χ^2 de Pearson requieren como supuestos:
- Dos variables nominales, cada una con un mínimo de dos valores,
- Estas variables deben ser temporalmente independientes; esto es, que no correspondan a un diseño de antes-después (este tipo de diseños se examinarán más adelante).

Para nuestro caso, tenemos una variable nominal politómica, tipología de familia, con cuatro valores: 1) familia sin padre ni madre, 2) familia monoparental paterna, 3) familia monoparental materna y 4) familia nuclear (con padre y madre). En este caso, cada progenitor se entiende como el padre biológico, o quien haga sus veces. Para la segunda variable, que define dos grupos de población, tenemos una variable nominal con dos valores: 1) ha tenido al menos un episodio de deserción de la escuela y 2) no ha tenido episodios de deserción de la escuela. Estas dos variables son diferentes e independientes, por lo que se cumplen los dos supuestos.

Sobre los tamaños de muestra, la prueba χ^2 de Pearson requiere:

- Un tamaño total de muestra mayor o igual a treinta individuos.
- Todas las celdas de la tabla de cruce deben tener una frecuencia de, al menos, cinco individuos.

Si estos supuestos de tamaño de la muestra no se cumplen, pueden seleccionarse, como alternativas, la prueba exacta de Fisher (disponible en SPSS y no disponible en JASP) o la prueba de razón de verosimilitud *(likelihood ratio*), disponible en los dos paquetes.

Para nuestro caso, observemos la tabla 60, de cruce entre las dos variables.

Tabla 60. Cruce entre tipo de familia y deserción parcial

	Se ha retirado del estudio?			
	Con deserción	**Sin deserción**	**Total**	
Sin padre ni madre	10	77	87	5,80 %
Monoparental paterna	8	39	47	3,14 %
Monoparental materna	54	355	409	27,28 %
Nuclear	81	875	956	63,78 %
Total	153	1346	1499	
	10,21 %	89,79 %		

Tal y como se observa, se cuenta con una muestra total de 1499 casos y ninguna de las celdas muestra una frecuencia de cinco casos o menos. En este sentido, también se cumple los supuestos de muestra para la aplicación de la prueba χ^2 de Pearson. En razón de que esta es la prueba más popular y conocida, examinaremos sus resultados.

- *Se examinan los resultados descriptivos.* La tabla 60 muestra que la proporción de individuos que han desertado de forma parcial en algún momento es de poco más del 10 %. Por otro lado, la tipología de familia más frecuente es la nuclear completa, que identifica a más del 63 % de los estudiantes, seguida por la familia monoparental de jefatura materna, con cerca del 27 % del total. Los otros tipos de familia tienen participaciones minoritarias.

A fin de comparar los dos grupos de individuos, la tabla se ha graficado en MS-Excel con barras apiladas al 100 % para mostrar la proporción de cada una de las tipologías de familia dentro del total de individuos según si han desertado en algún momento o no lo han hecho (figura 53).

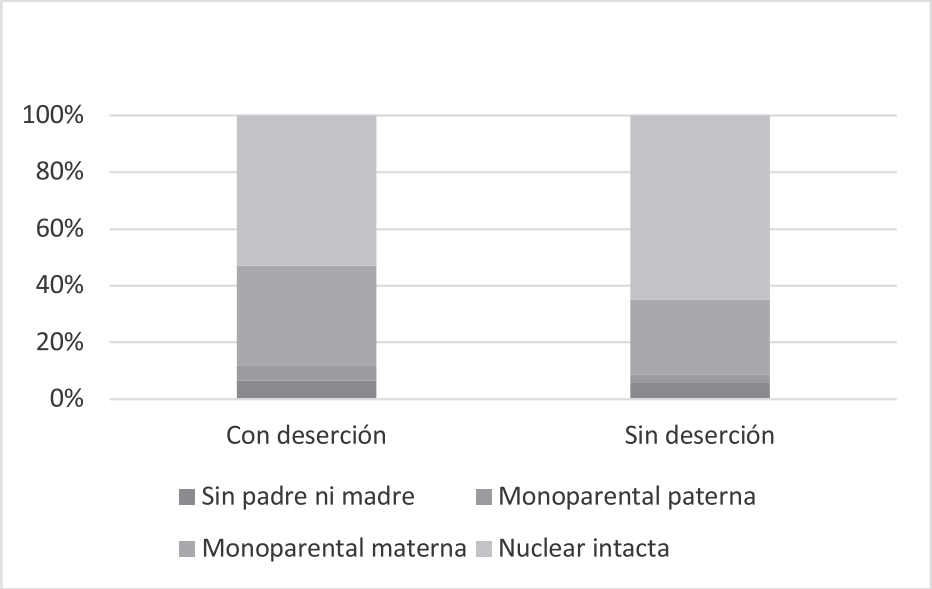


Figura 53. Tipo de familia por deserción parcial

Se tiene, entonces, que en los estudiantes que han mostrado al menos un episodio de deserción parcial, la proporción de familias nucleares es menor, mientras que la proporción de familias monoparentales de jefatura materna es mayor.

Se examinan los resultados de la prueba seleccionada

Para correr la prueba en el programa SPSS, debe seguirse esta secuencia: /Analizar/Estadísticos descriptivos/Tablas cruzadas… y allí se selecciona una de las variables en el campo "Filas" y la otra en el campo "Columnas", y se marca el botón "Estadísticos". En este menú se seleccionan dos casillas: "Chi-cuadrado" y "Phi y V de Cramer". El resultado de esta acción aparece en las tablas 61 y 62.

El resultado arrojado por el SPSS inicia, en la primera fila, con el valor del χ^2 de Pearson y su nivel de significación asociado. Como se observa, el resultado de la χ^2 de Pearson expone que las dos variables no son independientes. En otras palabras, las dos variables expresan dependencia y esta es estadísticamente significativa a un nivel de p=,021.

Tabla 61. Resultado de Chi-cuadrado en spss

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	9,681[b]	3	,021
Razón de verosimilitud	9,190	3	,027
Asociación lineal por lineal	5,337	1	,021
N de casos válidos	1499		

Tabla 62. Resultado de las medidas de tamaño del efecto en SPSS

		Valor	Significación aproximada
Nominal por Nominal	Phi	,080	,021
	V de Cramer	,080	,021
	Coeficiente de contingencia	,080	,021
N de casos válidos		1499	

En la segunda fila, se presenta el estadístico de razón de verosimilitud, bastante cercano al anterior y con una probabilidad asociada cercana, si bien es levemente mayor. En cualquier caso, ya sea que se interprete el χ^2 de Pearson o el estadístico de razón de verosimilitud, la conclusión es equivalente.

En la tercera fila, el SPSS presenta la "asociación lineal por lineal". Este estadístico no es apropiado para el tipo de datos nominales, que estamos manejando en este caso. Debe ser ignorado. En la tabla 61 se muestra el resultado relacionado con las medidas de tamaño del efecto apropiadas para variables nominales.

Como se evidencia, la tabla presenta el valor de Phi, el de la *V* de Cramer y el del coeficiente de contingencia. El Phi solo es apropiado para tablas 2x2 y el coeficiente de contingencia solo es apropiado para tablas de 5x5 o más. En la medida en que tenemos una tabla de 2x4, la única estimación válida es la *V* de Cramer, que, en este caso muestra un valor de ,080 con un *p*=,021.

Para la interpretación de la *V* de Cramer, necesitamos conocer los grados de libertad de la tabla. Para nuestro caso, los grados de libertad de la tabla serán

$$gl=(2-1)*(4-1)=3$$

Para gl=3, la tabla 60 indica que una *V* de Cramer de 0,08 corresponde a un tamaño del efecto que debe ser interpretado entre pequeño y mediano.

Se expresan los resultados

Para la expresión escrita de los resultados de la prueba χ^2 de Pearson, o sus similares, puede seguirse, en texto, el siguiente formato:

χ^2(<gl>)=<valor del Chi-cuadrado> p =<valor de p> V = <valor V de Cramer>

Aunque no es muy frecuente, es posible encontrar algunos textos en donde se incluyen los datos del tamaño de la muestra en este formato, de la siguiente forma:

χ^2(<gl>, N=<tamaño de muestra>)=<valor del Chi-cuadrado> p =
<valor de p> V = <valor V de Cramer>

En algunas ocasiones, la V de Cramer es denotada mediante el símbolo ϕ_C (Phi de Cramer) o V_C.

Atendiendo todas estas convenciones, la expresión de los resultados en texto podría quedar de la siguiente manera:

> *Los resultados indican que los estudiantes que muestran algún nivel de deserción parcial provienen, en mayor proporción, de familias monoparentales y en menor proporción de familias nucleares completas. Esta asociación muestra ser estadísticamente significativa en una prueba Chi-cuadrado de Pearson, si bien se indica un tamaño del efecto, medido por la V de Cramer, que puede ser valorado entre pequeño y mediano $\chi^2(3, N=1499)=9{,}68$ p=,021 $V = 0{,}08$.*

Pruebas para dos medidas apareadas (dos mediciones en la misma muestra)

Las pruebas para medidas relacionadas se aplican en situaciones de investigación en las cuales se han hecho dos mediciones de la misma variable en la misma muestra. Este tipo de situaciones es bastante frecuente en la investigación social y educativa, y responde a los que llamamos el diseño de medidas repetidas, o *diseño intrasujeto*. Este es el diseño que se utiliza para examinar cambios después de, por ejemplo, una intervención pedagógica.

Como lo hicimos en el caso de las pruebas sobre grupos independientes, ahora examinaremos las diferentes pruebas para muestras dependientes iniciando con el caso en que las variables tienen un nivel de medición de intervalo, y pasando después a los casos de variables ordinales y nominales.

El diagrama de la figura 54 muestra las condiciones y decisiones frente a cada prueba candidata a ser usada en un diseño intrasujeto. Cuando la variable es métrica y cumple el supuesto de normalidad, la elección adecuada es la prueba *t* de Student para medidas apareadas. En el caso en que no se cumpla el supuesto de normalidad, la prueba adecuada es una prueba no paramétrica: la W de Wilcoxon; la misma que se usaría si la variable dependiente tuviera un nivel ordinal de medida.

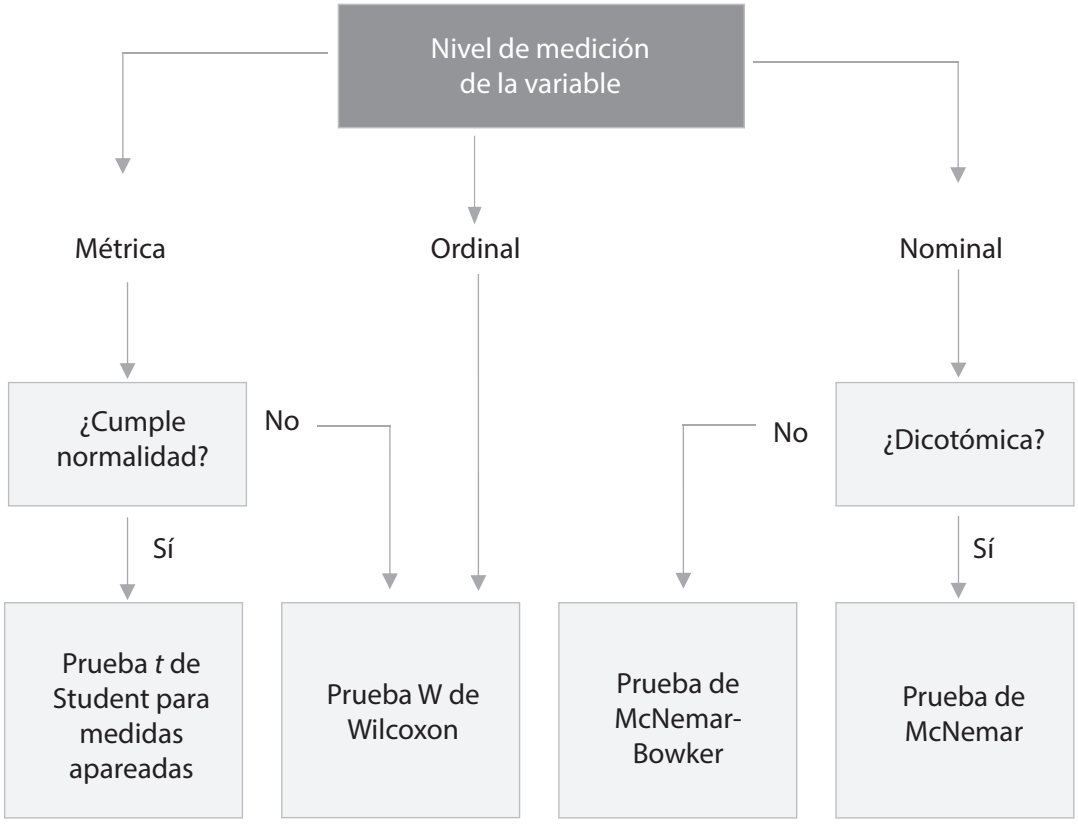


Figura 54. Pruebas para dos medidas apareadas

Ahora, si la variable dependiente tiene un nivel de medida estrictamente nominal, deben diferenciarse dos situaciones. Si la variable es dicotómica, la elección es la prueba de McNemar. Si, por el contrario, la variable es politómica, debemos proceder a una variación de la anterior, conocida como *prueba de McNemar-Bowker*.

Variable métrica: prueba t de Student para medidas apareadas/dependientes

Presentación

Cuando la variable tiene un nivel de medición de intervalo y tenemos un diseño con dos medidas repetidas, la candidata ideal es la *prueba t de Student para medidas apareadas* (también llamadas relacionadas, dependientes o correlacionadas).

La *prueba t de Student para medidas apareadas* es una prueba paramétrica que requiere de dos variables que representen dos momentos de medición, del mismo tipo, en las mismas personas. Como la prueba *t* de Student de grupos independientes, esta prueba produce un valor *t*, los grados de libertad (gl) y un nivel de significación asociado (*p*). El valor *t* representa la relación entre la diferencia de las medias en los dos momentos de medición y el error estándar de esta diferencia de medias.

Los supuestos de esta prueba son dos: 1) la normalidad de la diferencia entre los dos momentos de medición y 2) el que no debe haber valores atípicos muy significativos en la diferencia entre estas dos medidas. Esta prueba no requiere del cumplimiento del supuesto de homocedasticidad.

Como en el caso de la prueba para grupos independientes, esta prueba es robusta frente a violaciones moderadas del supuesto de normalidad. Este supuesto se verifica a través de la prueba de Shapiro-Wilk, que se encuentra incorporada en el mismo menú de la prueba en el paquete JASP, mientras que en el IBM-SPSS debe ser, primero, calculada la diferencia y después examinada su normalidad.

Tanto en el programa JASP como en el SPSS las salidas aportan información acerca del intervalo del confianza (IC), al 95 % o al 99 %, de la diferencia entre las dos medidas. En general, vale la pena reportar este intervalo.

Al respecto de las medidas de tamaño del efecto, la más conocida y utilizada en esta prueba *d* de Cohen, cuyos criterios de interpretación ya hemos expuesto antes. El cálculo de este estimador está incorporado en el paquete JASP en el mismo menú de la prueba, mientras que en el IBM-SPSS solo ofrece esta posibilidad en las últimas versiones.

Ejecutar la prueba t de Student para medidas apareadas

Para ejecutar la prueba *t* de Student para medidas apareadas en el programa JASP, puede seguirse el camino presentado en el recuadro 31. Para correr esta prueba en el IBM-SPSS puede seguirse la secuencia del recuadro 32.

Recuadro 31. Cómo ejecutar una prueba *t* para medidas apareadas en JASP

/T-Test/Classical/ Paired Samples T-Test.

En este punto, deben seleccionarse las parejas de variables apareadas (pueden ser varias) y pasarse a la lista "Variable pairs"

Test
- √ Student

Alt. Hypothesis
- √ Group 1 ≠ Group 2

Assumption Checks
- √ Normality

Additional Statistics
- √ Location parameter
  - √ Confidence interval [95,0 %]
- √ Effect Size
  - √ Confidence interval [95,0 %]
- √ Descriptives
- √ Descriptives plots
- Confidence interval [95,0 %]

Recuadro 32. Cómo ejecutar una prueba *t* para medidas apareadas en IBM-SPSS

/Analizar/Comparar medias/Prueba T para muestras relacionadas…
En este punto deben seleccionarse las parejas de variables (pueden ser varias) y pasarse a la lista "Variables emparejadas"
√ Estimar tamaños del efecto
Pulsar el botón "Aceptar"

El ejemplo

El ejemplo que trabajamos previamente con un profesor-investigador que diseña e implementa un programa de Matemáticas y quiere conocer su impacto mediante un diseño cuasiexperimental pretest/postest con grupo de control nos permite examinar la aplicación de pruebas sobre medias apareadas por cuanto se toman medidas de logro en Matemáticas antes de haber implementado programa y después de haberlo hecho.

Sin embargo, el diseño que se utilizó en este caso es un poco más complejo en la medida en que se dividió la muestra en dos grupos diferentes y solo uno de ellos cursó el programa de refuerzo en Matemáticas, mientras que el otro no lo cursó.

Interesa en este caso examinar si hubo un avance en cada uno de los grupos por separado. Por esta razón, examinaremos una prueba *t* para medias apareadas en el grupo experimental, en el que se implementó el programa, y otra diferente en el grupo de control, en el que no se experimentó. Tenemos, entonces, dos situaciones en las que aplicaremos una prueba *t* para medidas apareadas. Desarrollaremos estas dos situaciones de forma paralela.

Planteamiento de las hipótesis

El planteamiento de las hipótesis no ofrece grandes novedades. Iniciando con el examen de diferencias en el grupo experimental:

> *Hipótesis nula. No hay diferencias entre las medias de las pruebas de Matemáticas en el pretest y postest en el grupo experimental.*
>
> *Hipótesis alternativa. Existen diferencias en las medias de las pruebas de Matemáticas en el pretest y en el postest en el grupo experimental.*

Aunque la forma de esta hipótesis es igual para las dos pruebas *t* que examinaremos, esperamos encontrar que, en el grupo experimental, se pueda rechazar la hipótesis nula, esto es, que encontramos diferencias significativas entre las medias del pretest y del postest. En realidad esperamos más que esto: quisiéramos encontrar que las medias de las pruebas del postest sean significativamente mayores que las del pretest. En ese sentido, tendríamos una hipótesis unilateral. Sin embargo, tal y como lo hemos justificado antes, la formulación bilateral de la hipótesis es suficiente para nosotros.

El planteamiento de las hipótesis para el grupo de control es equivalente:

> *Hipótesis nula. No hay diferencias entre las medias de las pruebas de Matemáticas en el pretest y postest en el grupo de control.*
>
> *Hipótesis alternativa. Existen diferencias en las medias de las pruebas de Matemáticas en el pretest y en el postest en el grupo de control.*

En relación con lo que esperamos encontrar en el grupo de control la situación es más complicada. Si no pudiéramos rechazar la hipótesis nula en el grupo de control, y si lo hiciéramos en el grupo experimental, la interpretación de los resultados sería sencilla y la efectividad del programa de Matemáticas quedaría verificada. Sin embargo, podríamos esperar, razonablemente, un incremento en las medias entre las dos pruebas en el grupo de control.

Las razones de este hipotético incremento en el grupo de control son múltiples. Por un lado, podría considerarse que en el postest, los estudiantes ya llegan con experiencia acerca de la naturaleza de la prueba y de las preguntas, lo cual podría incrementar su desempeño. Por otro lado, es posible que algunos estudiantes del grupo de control hayan utilizado el tiempo presente entre las dos pruebas para mejorar sus competencias en el área. Esperaríamos, eso sí, encontrar que los niveles de avance en el grupo de control sean menores que los observados en el grupo experimental, para lo cual las medidas de tamaño del efecto pueden ser útiles. De otra forma, tendríamos que concluir que el programa de Matemáticas no contribuye de forma especialmente notable.

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba*. La prueba *t* para medias dependientes supone el cumplimiento del supuesto de normalidad para la diferencia de medias. Las tablas 63 y 64 muestran los resultados de las pruebas de normalidad según se presentan en el programa JASP.

Tabla 63. Grupo experimental. Prueba de normalidad para la diferencia de medias entre el pretest y el postest (Shapiro-Wilk)

			W	p
premat	-	posmat	0,971	,681

Nota: resultados significativos sugieren una desviación de la normalidad.

Tabla 64. Grupo de control. Prueba de normalidad para la diferencia de medias entre el pretest y el postest

			W	p
premat	-	posmat	0,964	,492

Nota: resultados significativos sugieren una desviación de la normalidad.

Como se observa en las tablas 63 y 64, el supuesto de normalidad para la diferencia de medias se cumple en los dos casos. Las pruebas de normalidad de Shapiro-Wilk muestran que la diferencia de medias no dista de la curva normal, ni en el grupo experimental W=0,971 p=,681, ni en el de control W=0,964 p=,492. En estas condiciones, podemos aplicar la prueba t de Student para muestras apareadas.

- *Se examinan los resultados descriptivos.* La tabla 65 y la figura 55 muestran la comparación de las medias del pretest y del postest para los grupos experimental y de control.

Tabla 65. Medias, desviaciones estándar y errores estándar de pretest y postest en los grupos experimental y de control

Grupo	Prueba	N	Media	DE	EE
Grupo experimental	Pretest	25	78,28	20,88	4,18
	Postest	25	102,32	31,94	6,39
Grupo de control	Pretest	25	78,37	18,98	3,80
	Postest	25	81,32	26,13	5,23

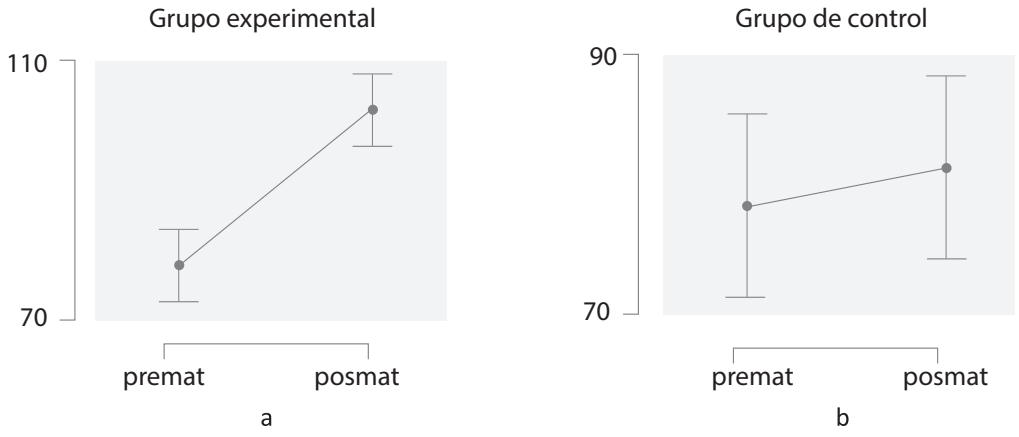


Figura 55. Diferencias entre pretest y postest en los grupos experimental y control

En la tabla 64 se evidencia que las medias del postest son siempre mayores que las del pretest para los dos grupos. Sin embargo, es claro que la diferencia entre los dos valores es mucho más acentuada en el grupo experimental.

Tal y como se observa, para los dos grupos examinados las medias del postest de la prueba de Matemáticas son mayores que las medias presentadas en el pretest en la misma prueba. Sin embargo, la simple inspección visual de las gráficas indica que la diferencia entre el pretest y el postest es bastante mayor en el grupo experimental que en el de control

Se examinan los resultados de la prueba

Las tablas muestran los resultados arrojados por el programa JASP para cada uno de los dos grupos (tabla 66). Iniciando por el grupo experimental (tabla 67), se presentan diferencias significativas

entre las dos medias a niveles inferiores a ,001. Consecuentemente, el intervalo de confianza, al 95 %, de la diferencia de las medias no contiene el valor 0, lo que confirma el rechazo de la hipótesis nula para este grupo.

Tabla 66. Resultados de la prueba t para medias apareadas entre pretest y postest en el grupo experimental

								IC 95 % para la diferencia		
			t	gl	p	Diferencia media	EE de la diferencia	Bajo	Alto	d de Cohen
posmat	-	premat	6,192	24	<,001	24,05	3,883	16,03	32,06	1,238

Tabla 67. Resultados de la prueba t para medias apareadas entre pretest y postest en el grupo de control

								IC 95 % para la diferencia		
			t	gl	p	Diferencia media	EE de la diferencia	Bajo	Alto	d de Cohen
posmat	-	premat	0,605	24	0,551	2,944	4,866	-7,10	12,99	0,121

Por otro lado, el examen de la prueba *t* sobre medias apareadas en el grupo de control muestra que el valor de la significación de la diferencias es *p*=,551 es bastante mayor que el nivel aceptado, lo cual indica que no hay diferencias estadísticamente significativas entre las dos medias al nivel de 0,05. Consecuentemente, el intervalo de confianza para la diferencia de medias contiene el valor "0", lo que confirma la imposibilidad de rechazar la hipótesis nula. El valor de la *d* de Cohen para la estimación del tamaño del efecto es muy bajo, ubicándose en el rango que se identifica con ausencia de efecto.

Como se observa, los valores *t* arrojados por las pruebas muestran ser positivos, así como las diferencias medias y los tamaños del efecto. Este signo solo indica la dirección del cambio. En este caso, al incluir las dos variables, introducimos el postest (posmat) antes del pretest (premat). En la medida en que los postest fueron mayores que los pretest en los dos casos, la diferencia postest-pretest fue positiva en los dos casos. Si los hubiéramos introducido en el orden contrario, la diferencia habría sido negativa, lo que también habría afectado el signo de la *d* de Cohen. En general, el signo de los valores *t* y *d* solo representa la dirección del cambio.

Se expresan los resultados

Para la expresión de los resultados de la prueba *t* para medias apareadas, se sigue el mismo formato que ya habíamos visto antes en la prueba *t* para grupos independientes:

t(<gl>)= <valor t> *p* = <valor p> *d* = <d de Cohen> IC 95 % [<LI> , <LS>]

En donde:

gl: grados de libertad
LI: límite inferior del intervalo de confianza
LS: límite superior del intervalo de confianza
Los valores entre los signos < y > son los valores arrojados por la prueba.

Utilizando este formato, podríamos expresar los resultados de la siguiente forma como sigue:

> *Las medias (con las deviaciones estándar entre paréntesis) para el grupo experimental fueron 78,28 (20,88) para el pretest y 102,32 (31,95) para el postest; la diferencia entre estas dos medias es significativa y el tamaño del efecto muestra ser grande en este grupo t(24) = 6,19 p < ,001 d = 1,24 IC 95 % [16,03, 32,06]. En contraste con ello, las medias para el grupo control fueron 78,37 (18,98) para el pretest y 81,32 (26,12) para el postest y la diferencia entre las dos medias no muestra ser significativa en el nivel de ,05, mientras que el tamaño del efecto es prácticamente nulo t(24) = -0,61 p = ,551 (NS) d = 0,121 IC 95 % [-7,10, 12,99].*

Como ya lo hemos mencionado antes, estos resultados pueden ser expresados en tablas, especialmente si son muchos. Para el caso de solo dos pruebas, expresar los resultados en el texto es lo adecuado.

Variable ordinal: prueba de los signos de Wilcoxon

Presentación

Existe una prueba no paramétrica equivalente a la prueba *t* para muestras apareadas: la prueba de los signos de Wilcoxon, o *prueba de Wilcoxon de los rangos con signo*. Se utiliza en diseños antes-después, o intrasujeto, en los casos en los que tenemos variables continuas cuya diferencia no se encuentra normalmente distribuida, presenta valores atípicos muy significativos o tiene un nivel de medida ordinal.

La prueba de Wilcoxon de los rangos con signo es una prueba no paramétrica, que, como la prueba U de Mann-Whitney, calcula las diferencias entre las dos variables apareadas teniendo en cuenta el signo y la magnitud de las diferencias. Produce un estadístico (*W*) y un valor de la significación asociado con este estadístico.

Como en el caso de las otras pruebas no paramétricas, la prueba de Wilcoxon de los rangos con signo no requiere del cumplimiento del supuesto de normalidad. La única condición para su aplicación es la presencia de dos variables que representen dos momentos de medición, del mismo constructo en los mismos sujetos, con un nivel de medida, al menos, ordinal.

En general, en todos los programas que utilizamos se presenta información de estadísticas descriptivas de las dos medidas: medias, desviaciones estándar y errores estándar, que suelen ser suficientes para examinar la dirección y el sentido de las diferencias entre las dos medidas.

Por otro lado, las salidas de los dos programas que utilizamos difieren en gran medida para esta prueba. Como en el caso de la prueba U de Mann-Whitney, el programa JASP reporta un estimador de la diferencia mediada entre los rangos conocido como el estimador de Hodges-Lehmann,

con su correspondiente intervalo de confianza. En el caso del IBM-SPSS, se presenta una tabla con rangos promedio y sumas de rangos.

Respecto de las medidas de tamaño del efecto en esta prueba, también se presentan diferencias entre los programas. Como en el caso de la prueba U de Mann-Whitney, el programa JASP reporta la correlación rango biserial (r_{rankb}) como la medida de tamaño del efecto apropiada para esta prueba; para su interpretación, puede consultarse la tabla 68.

Tabla 68. Interpretación de los valores de r o r_b como medidas de tamaño del efecto

r	Interpretación según Cohen (1988)
r<0,1	Sin efecto
0,1<r<0,3	Efecto pequeño
0,3<r<0,5	Efecto medio
0,5 < r	Efecto grande

El IBM-SPSS no reporta ningún indicador de tamaño del efecto en esta prueba. En la página www.psychometrica.de, es posible calcular otra medida de tamaño del efecto (eta cuadrado) para esta prueba y convertirla en otras (*d* de Cohen, por ejemplo).

Ejecutar la prueba de los signos de Wilcoxon

Para ejecutar la prueba de los signos de Wilcoxon en el programa JASP, pueden tomarse las opciones del recuadro 33. Para correr la prueba en el IBM-SPSS, debe procederse a través del menú /Analizar/Pruebas no paramétricas (recuadro 34).

Recuadro 33. Cómo ejecutar una prueba de Wilcoxon en JASP

/T-Test/Classical/ Paired Samples T-Test.
En este punto, deben seleccionarse las parejas de variables apareadas (pueden ser muchas) y pasarse a la lista "Variable pairs"
Test
√ Wilcoxon signed-rank
Alt. Hypothesis
√ Group 1 ≠ Group 2
Additional Statistics
√ Location parameter
√ Confidence interval [95,0 %]
√ Effect Size
√ Confidence interval [95,0 %]
√ Descriptives
√ Descriptives plots
Confidence interval [95,0 %]

Recuadro 34. Cómo ejecutar una prueba de Wilcoxon en *IBM-SPSS*

/Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/2 muestras relacionadas…
En este punto deben pasarse una, o varias, parejas de variables a la lista "Contrastar pares"
√ Wilcoxon
Pulsar el botón "Aceptar"

El ejemplo

Presentaremos el uso de esta prueba de los signos del Wilcoxon a partir de nuestro ejemplo relacionado con las actitudes que presentan los estudiantes de forma previa y posterior a la aplicación de un programa de refuerzo en dos grupos por separado: el experimental y el de control. En particular examinaremos 1) las diferencias entre las actitudes frente a la asignatura de Matemáticas en el grupo experimental de forma previa y posterior a la aplicación del programa, y 2) las diferencias entre las actitudes frente a la asignatura de Matemáticas en el grupo control de forma previa y posterior a la aplicación del programa.

Planteamiento de las hipótesis

Ya en este punto no debemos tener dificultades en el planteamiento de las hipótesis. Iniciamos con las diferencias en las actitudes frente a la asignatura de Matemáticas en el grupo experimental.

> *Hipótesis nula. No hay diferencias entre las medianas de las actitudes frente a las asignaturas de Matemáticas en el pretest y postest en el grupo experimental.*
>
> *Hipótesis alternativa. Existen diferencias entre las medianas de las actitudes frente a las asignaturas de Matemáticas en el pretest y postest en el grupo experimental.*

En cuanto a las diferencias entre las actitudes presentadas en el grupo de control, las formas son idénticas. Solo difiere la mención del grupo de control.

> *Hipótesis nula. No hay diferencias entre las medianas de las actitudes frente a las asignaturas de Matemáticas en el pretest y postest en el grupo de control.*
>
> *Hipótesis alternativa. Existen diferencias entre las medianas de las actitudes frente a las asignaturas de Matemáticas en el pretest y postest en el grupo de control.*

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* La prueba de los signos de Wilcoxon solo requiere dos supuestos. Primero, que la variable que se compara sea la misma, en dos momentos diferentes de medición. Este presupuesto se cumple, por la naturaleza del diseño. Segundo, que la variable que se compara tenga, al menos, un nivel de medida ordinal. Este presupuesto se cumple. Para nuestro caso, la medición de las actitudes tiene un nivel de medida ordinal, de cuatro puntos: 1) negativa, 2) neutra, 3) positiva y 4) muy positiva.

Dado que todos los presupuestos de la prueba de los signos de Wilcoxon se cumplen, podemos aplicar y examinar los resultados de las dos pruebas. Primero, examinaremos la naturaleza de los cambios entre las dos mediciones en cada uno de los grupos.

- *Se examinan los resultados descriptivos.* La tabla presenta los datos del cruce entre las variables de actitud en el postest y el pretest para cada uno de los grupos. Como se observa, en el grupo experimental la moda en el pretest corresponde a una actitud neutra (f=12) mientras que en el postest la moda es una actitud positiva (f=11). Esto pareciera sugerir que la actitud mejoró entre el pre y el postest en ese grupo. Por su parte, en el grupo de control la moda de la actitud en el pretest es neutra y positiva (f=7 en los dos casos), mientras que en el postest es positiva (f=11). Esto pareciera también sugerir que la actitud entre las dos mediciones mejoró en alguna medida. Sin embargo, esta apreciación es claramente incompleta. Para una apreciación más general debemos examinar la representación gráfica de estos mismos datos (tabla 69).

Tabla 69. Cruce entre las variables de actitud en el postest y el pretest para cada uno de los grupos

Grupo	Prueba	Actitud hacia la matemática			MPositiva	Total
		Negativa	Neutra	Positiva		
Experimental	Pretest	5	12	7	1	25
	Postest	0	5	11	9	25
Control	Pretest	5	7	7	6	25
	Postest	3	7	11	4	25

La figura 56 muestra la proporción de los diferentes valores de la actitud, en la forma de una gráfica de barras apiladas al 100 %.

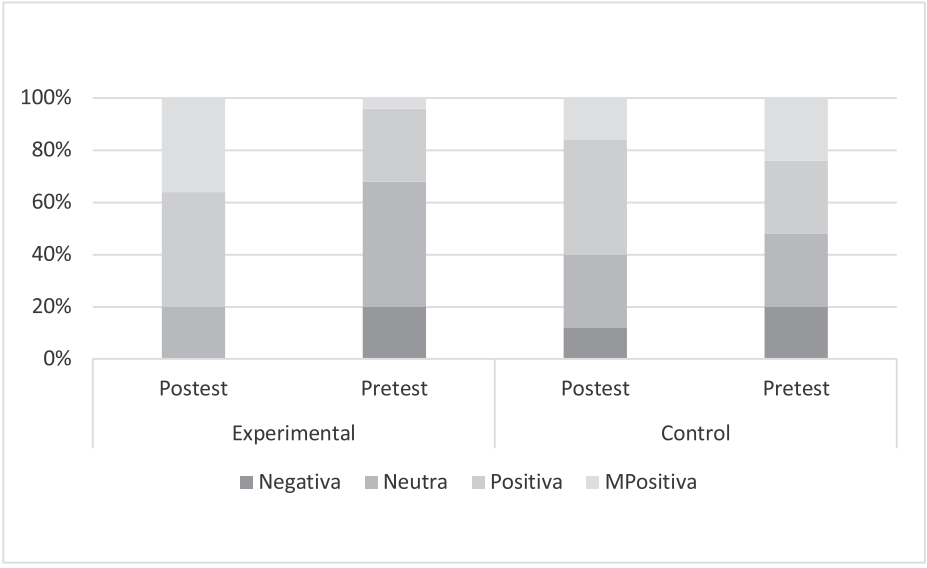


Figura 56. Actitudes postest y pretest en los grupos experimental y de control

Iniciando en el grupo experimental, la inspección visual muestra que las proporciones de actitudes negativas y neutras disminuyeron, mientras que las proporciones de actitudes positivas y muy positivas aumentaron entre las dos medidas. Esto confirma la idea de que las actitudes cambiaron, en el sentido de hacerse mucho más positivas para el postest.

Por el lado del grupo de control, se observa para el postest una disminución de las actitudes negativas, niveles similares de actitudes neutras, un aumento de las actitudes positivas y una cierta disminución de las actitudes muy positivas. Visto este panorama de forma más completa, es difícil saber si las actitudes mejoraron o no lo hicieron en el paso del pretest al postest.

A esta misma conclusión se llega si observamos las gráficas de medias y errores estándar que presenta el programa JASP en la sección de pruebas apareadas. Tal y como se observa en la figura 57, se presenta un incremento importante en la actitud positiva en el grupo experimental (figura 57a) mientras que para el grupo de control (figura 57b), aunque se observa un incremento en la actitud positiva, este no parece tan marcado.

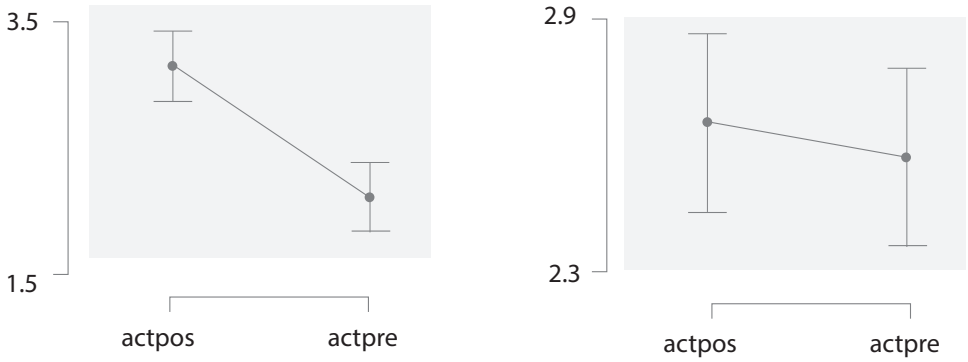


Figura 57. Medias y errores estándar de los puntajes de actitud en el pretest y postest

Nota: a) en el grupo experimental; b) en el grupo de control.

La confluencia de estos dos resultados haría pensar que, aunque no resulte formalmente apropiado verificar el cambio en las actitudes por el cambio en las medias entre los dos momentos de medición, esta puede ser la forma más sencilla y rápida de hacerlo.

En general, cabe esperar diferencias en los dos grupos en el sentido de que en el postest los estudiantes muestran actitudes más positivas que en el pretest, y muy especialmente en el grupo experimental. La medida en que este cambio resulte estadísticamente significativo será lo que examinaremos al correr la prueba e interpretar los resultados.

Se examinan los resultados de la prueba

Como en los casos anteriores, hemos solicitado al programa JASP las pruebas de Wilcoxon para las dos comparaciones que estamos haciendo, con el estimador de localización, su intervalo de confianza asociado y una medida del tamaño del efecto. Los resultados de las dos pruebas examinadas, tal y como son presentados por el programa JASP se muestran en las tablas 70 y 71 para el grupo experimental y de control, respectivamente.

Tabla 70. Grupo experimental. Resultado de las diferencias entre postest y pretest de actitudes en la prueba de Wilcoxon

					IC 95 % para estimado de Hodges-Lehmann		
		W	p	Estimado de Hodges-Lehmann	Bajo	Alto	Correlacion rango-biserial
actpos	- actpre	136,0	< ,001	1,500	1,000	2,000	1,000

Note. Test de los signos de Wilcoxon.

Tabla 71. Grupo de control. Resultado de las diferencias entre postest y pretest de actitudes en la prueba de Wilcoxon

					IC 95 % para estimado de Hodges-Lehmann		
		W	p	Estimado de Hodges-Lehmann	Bajo	Alto	Correlacion rango-biserial
actpos	- actpre	13,50	0,589	0,500	-1,000	1,500	0,286

Iniciando por la comparaciones en el grupo experimental, los resultados indican que la diferencia mediana es significativamente diferente de cero (p<,001). Correspondiente con ello, el estimador Hodges-Lehmann para la diferencia mediana entre las dos pruebas es de 1,50, y su intervalo de confianza asociado, al 95 %, no incluye el valor "0". La estimación del efecto, a través de la correlación rango biserial, indica un valor de 1, que señala una estimación de tamaño de efecto grande.

Por el lado del grupo de control, la prueba de Wilcoxon señala que la diferencia mediana no es significativa al nivel de 0,05 (p=0,589>0,050), lo cual se confirma en el intervalo al 95 % de confianza que incluye el valor "0". La correlación de rango biserial, usada en este caso como estimación de tamaño del efecto, señala un valor de efecto que debe ser considerado pequeño.

Se expresan los resultados

Los resultados pueden ser expresados en un texto de la siguiente forma:

> *Las diferencias entre el pretest y el postest de actitudes fueron examinadas, en cada grupo, con una prueba de los rangos con signo de Wilcoxon. Iniciando con el grupo experimental, los resultados muestran que las actitudes en el postest son significativamente más positivas que las presentadas en el pretest, y que el efecto del programa en el mejoramiento de las actitudes es grande W=136,00 p<,001 r_b=1,00. En contraste con ello, las diferencias entre el pretest y el postest en las actitudes en el grupo de control no son estadísticamente significativas y el efecto es de tamaño pequeño W=13,50 p=,589 (ns) r_b=,286.*

Variable nominal: pruebas de McNemar y McNemar-Bowker

Presentación

Cuando tenemos una situación de medidas apareadas (muestras relacionadas) en la que la variable es de naturaleza nominal, no podemos utilizar pruebas *t* de Student para medidas apareadas, ni pruebas de Wilcoxon. En este caso, se debe utilizar la *prueba de McNemar*, para variables dicotómicas, o una de sus extensiones, conocida como la *prueba de McNemar-Bowker*, para variables nominales politómicas.

Estas pruebas se basan en la distribución Chi-cuadrado y examinan la tabla de cruce de la medición previa con la medición posterior. En la medida en que estamos analizando diseños intrasujeto, en los que tenemos dos mediciones iguales de la misma variable en dos momentos diferentes, la tabla de cruce de estas dos mediciones siempre es una tabla cuadrada de orden m*m, que podríamos representar como:

$$(a_{ij})\ i,j=1,\ldots,m$$

En esta situación, en la diagonal principal (a_{ii}) se registran los individuos que no cambiaron de una medición a otra, mientras que por fuera de esta diagonal ($a_{ij,}\ i\neq j$) están los individuos que sí cambiaron su respuesta en la segunda medición. Estas pruebas comparan el número de individuos que no cambiaron con el número de los que sí cambiaron, para producir un estadístico específico (*M*, para la prueba de McNemar y *B* para la prueba de McNemar-Bowker) con un nivel de significación asociado en cada una de estas pruebas (*p*).

Estas pruebas son pruebas no paramétricas de homogeneidad y simetría. En la prueba de McNemar, solo se tiene el supuesto de dos mediciones, antes y después, de una variable *dicotómica*. En la prueba de McNemar-Bowker se admite que esta variable sea nominal politómica.

Al respecto de las medidas de tamaño del efecto, debe mencionarse que no hay un consenso sobre las medidas adecuadas a cada una de estas pruebas. Los paquetes estadísticos que las presentan y procesan no incluyen ninguna medida al respecto.

Ejecutar las pruebas de McNemar y McNemar-Bowker

Las pruebas de McNemar y McNemar-Bowker están incluidas en el paquete IBM-SPSS en la parte dedicada a las tablas de contingencia. Siga el procedimiento del recuadro 35.

Recuadro 35. Ejecutar la prueba de McNemar en IBM-SPSS

/Analizar/Estadísticos descriptivos/Tablas cruzadas….
Allí se seleccionan, una variable en filas y otra en columnas (el orden no tiene importancia)
    En el botón "Estadísticos"
        √ "McNemar"
    Se oprime "Continuar"
Se oprime "Aceptar".

En este punto, el programa presentará los resultados de la prueba adecuada a la cantidad de valores de la variable incluida. En el paquete JASP no se incluyen las pruebas de McNemar ni de McNemar Bowker.

El ejemplo: efectos de un programa para favorecer el empleo

Una conocida ONG internacional promueve y financia un programa educativo cuyo potencial efecto es favorecer las condiciones de ocupación de los estudiantes que los tomen. Para ello, mide el estado de ocupación de una muestra antes de iniciar al programa, aplica el programa en la muestra de estudiantes y vuelve a examinar los resultados del estado de ocupación después de dos meses de terminado el programa.

El *estado de ocupación* es una variable nominal politómica que presenta las siguientes categorías:

- Inactivo. Son asistentes que no laboran ni desean hacerlo.
- Desempleado. Corresponde a individuos que no se encuentran laborando y quisieran hacerlo, por lo que se encuentran en búsqueda de empleo.
- Independiente. Corresponde a individuos que realizan trabajos varios de carácter temporal y más o menos informal.
- Empleado. Corresponde a individuos empleados formalmente a término indefinido en una empresa legalmente constituida.
- Emprendedor. Corresponde a individuos que son propietarios de la empresa en la que laboran.

El programa educativo fue diseñado para favorecer la constitución de emprendimientos. Se aplicó, por parte de una universidad local, en una muestra de 140 personas. Tenemos en este caso una variable nominal que se examina antes, y después, de una intervención educativa. Interesa conocer los efectos del programa en el estado de ocupación de los asistentes.[7]

Planteamiento de las hipótesis

Las hipótesis podrían ser planteadas de la siguiente forma:

> *Hipótesis nula (H_0). No existen diferencias entre las condiciones de empleo antes de tomar el programa y las mismas condiciones tres meses después de terminado este.*
>
> *Hipótesis alternativa (H_1). Existen diferencias entre las condiciones de empleo antes de tomar el programa y tres meses después de concluido este.*

7 Aunque el ejemplo presentado es genuino y corresponde a una situación real de consultoría, los datos presentados son ficticios.

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba*. Las pruebas de McNemar y McNemar-Bowker no tienen más supuestos que los determinados por el diseño (diseño intrasujeto con dos mediciones: antes y después) y por el nivel de medida de la variable (nominal monotónica en el caso de la prueba de McNemar y nominal politómica en el caso de la prueba de McNemar-Bowker).
- Se examinan resultados descriptivos. La tabla 72 muestra el cruce entre las dos mediciones, antes (filas) y después (columnas).

Tabla 72. Tabla de cruce entre las condiciones de empleo antes de tomar el programa y después de hacerlo

		Después					Total
		Inactivo	Desempl.	Independ.	Empleado	Emprend.	
Antes	Inactivo	5	8	8	8	8	37
	Desempleado	4	6	9	9	7	35
	Independiente	2	2	4	7	11	26
	Empleado	1	3	4	10	3	21
	Emprendedor	3	5	1	3	9	21
	Total	15	24	26	37	38	140

Como se observa, el total de casos en la diagonal principal no es muy alto (34 casos, dentro de un total de 140 casos). Esto nos induce a pensar un cambio considerable entre el estado de ocupación previo y el posterior al programa.

En la medida en que nos interesa, para este caso, examinar el efecto específico del programa en la constitución de emprendimientos examinaremos, por separado, las transformaciones de cada una de las categorías examinadas en la de emprendimiento a través de la aplicación de pruebas de McNemar. Las siguientes son las tablas de cruce entre cada categoría (inactivo, desempleado, independiente y empleado) y la categoría de "emprendedor" (tabla 73).

Una rápida inspección visual de estas tablas mostraría que la condición más estable, y por tanto con menos cambios, es la condición del empleado (diez de ellos siguieron en el mismo estado), mientras que la más cambiante puede ser la del independiente (solo cuatro de ellos siguieron siéndolo después del programa).

Tabla 73. Tablas de cruce entre la categoría "emprendedor" y las categorías inactivo, desempleado, independiente y empleado

		Después		Total
		Inactivo	Emprendedor	
Antes	Inactivo	5	8	13
	Emprendedor	3	9	12
	Total	8	17	25
Antes	Desempleado	6	7	13
	Emprendedor	5	9	14
	Total	11	16	27
Antes	Independiente	4	11	15
	Emprendedor	1	9	10
	Total	5	20	25
Antes	Empleado	10	3	13
	Emprendedor	3	9	12
	Total	13	12	25

Se examinan los resultados de la prueba

La tabla 74 presenta los resultados de la prueba de McNemar-Bowker, tal y como son expuestos por el SPSS para el cruce con todas las categorías de ocupación presentes. Como se observa, la prueba indica diferencias significativas (p=,001) entre el estado de empleo previo y el estado de empleo posterior. La tabla 75 consolida todos los resultados obtenidos de las tablas de cruce entre cada una de las categorías presentes y el emprendimiento.

Tabla 74. Pruebas de Chi-cuadrado

Prueba de McNemar-Bowker	p	Casos válidos
Inactivo vs. Emprendedor	,227	25
Desempleado vs. Emprendedor	,774	27
Independiente vs. Emprendedor	,006	25
Empleado vs. Emprendedor	1,00	25

Tabla 75. Pruebas de Chi-cuadrado

		Valor	**Significación exacta (bilateral)**
	Prueba de McNemar		,227[a]
	N.º = de casos válidos	25	
Desempleado vs. emprendedor	**Pruebas de Chi-cuadrado**		
		Valor	**Significación exacta (bilateral)**
	Prueba de McNemar		,774[a]
	N.º = de casos válidos	27	
Independiente vs. emprendedor	**Pruebas de Chi-cuadrado**		
		Valor	**Significación exacta (bilateral)**
	Prueba de McNemar		,006[a]
	N.º = de casos válidos	25	
Empleado vs. emprendedor	**Pruebas de Chi-cuadrado**		
		Valor	**Significación exacta (bilateral)**
	Prueba de McNemar		1,000[a]
	N.º = de casos válidos	25	

a. Distribución binomial utilizada.

De este modo, las diferencias más significativas se presentan en el paso del independiente al emprendedor. De quince sujetos que eran independientes en la primera medición, once se presentan como emprendedores en la segunda medición. En las otras categorías la diferencia no es significativa.

Se expresan los resultados

Los resultados pueden ser expresados de la siguiente forma:

> *La prueba de McNemar-Bowker evidencia diferencias estadísticamente significativas entre las condiciones de ocupación presentes antes y después del programa B(10)=29,59 p=,001. El examen de las transformaciones de cada una de las categorías de ocupación en la de emprendimiento, realizada a través de sucesivas pruebas de McNemar, indica que la única categoría con diferencias significativas es la de independiente p=,006; las otras (inactivo, desempleado y empleado) no muestran cambios significativos hacia el emprendimiento p=,227, p=,774 y p=1,000, respectivamente. Puede concluirse que el programa indica cambios significativos hacia la constitución de emprendimientos en los participantes, específicamente si ellos se encuentran previamente en la categoría de ocupación independiente.*

En la medida en que estas pruebas no son muy utilizadas, puede ser más útil y claro referirse a los estadísticos asociados con estas pruebas como χ^2 de McNemar o χ^2 de McNemar-Bowker.

Capítulo 11

Pruebas de diferencias entre *k* medidas (tres o más)

En el capítulo anterior presentamos las pruebas más importantes para la comparación de dos medidas. En este, haremos una extensión de lo presentado para el caso en el que queremos comparar tres medidas o más.

Al igual que lo hicimos en el capítulo anterior, distinguimos en este caso dos tipos de comparación cualitativamente diferentes: la comparación entre *k* grupos independientes y la comparación entre *k* medidas apareadas. La diferencia ya debe ser clara: en un caso estamos interesados en la comparación entre diferentes grupos, y en el otro, en la comparación entre diferentes mediciones que hemos hecho, en las mismas personas, a lo largo del tiempo.

En cada uno de estos casos, la selección de la prueba inicia con la consideración del nivel de medida que tenemos; métrica, ordinal o nominal. Una vez sabemos esto, se trata de verificar el cumplimiento de los supuestos y, dependiendo de este resultado, proceder a la aplicación de la prueba y al examen y expresión de resultados y conclusiones.

Iniciaremos con el caso de la comparación entre tres, o más, grupos independientes.

Pruebas para *k* grupos independientes

La selección de la prueba adecuada para la comparación de tres o más grupos independientes depende, en primera instancia, del nivel de medición de la variable dependiente y, en segunda, del grado de cumplimiento de los supuestos de la prueba misma.

Ya hemos dicho en repetidas ocasiones que, de poder elegir, siempre preferiríamos la selección de una prueba paramétrica que, en este caso, corresponde al análisis de varianza de una vía (Anova *one way*). Esta prueba puede ser utilizada en el caso en el que tengamos una variable dependiente métrica intervalar y cumplamos los supuestos usuales para las pruebas paramétricas: la normalidad de la variable dependiente y la homogeneidad de varianzas.

En el caso del incumplimiento del supuesto de homogeneidad de varianzas, el Anova de una vía ofrece varias posibilidades de corrección (correcciones de Brown-Forsythe y Welch), por lo que esta situación no resulta insalvable. Ahora, en el caso de un incumplimiento muy serio del supuesto

de normalidad, y a pesar de que sabemos que el Anova de una vía es bastante robusto frente a desviaciones moderadas de este supuesto, en algunos casos esto es imposible de conseguir. En estos casos, deberemos proceder al uso de la alternativa no paramétrica al Anova: la prueba H de Kruskal-Wallis. Sin embargo, la prueba H supone el cumplimiento del supuesto de homocedasticidad, especialmente en muestras muy pequeñas, por lo que si tenemos una variable dependiente métrica, que no cumple con los supuestos de normalidad ni homocedasticidad y la muestra es muy pequeña, es preferible la opción de aplicar el Anova con corrección de Welch. Para los casos en que tenemos una variable dependiente ordinal la prueba H es la única opción adecuada, si bien sería adecuado examinar el supuesto de homogeneidad de varianzas, especialmente en muestras muy pequeñas ($n<30$).

Finalmente, para el caso de que la variable dependiente tenga un nivel de medida nominal, volvemos a la aplicación de la prueba Chi-cuadrado en tablas de contingencia, que ya habíamos examinado en el capítulo anterior, si bien ahora la variable independiente tendrá tres valores o más.

Una vez hemos constatado que existen diferencias globales entre las medias o las distribuciones de los diferentes grupos, deberemos examinar cuáles son los grupos específicos en los que se presentan estas diferencias. El examen de estas diferencias se hace a través de pruebas *post hoc* (del latín: "después de esto").

Es importante dejar claro que las pruebas *post hoc* solo deben ser examinadas en el caso en el que la prueba global haya indicado diferencias significativas entre los grupos. De otra forma, incurrimos en un aumento muy importante de la probabilidad de errores de tipo I, por lo que perderíamos confianza en los resultados de la prueba.

La prueba *post hoc* apropiada dependerá de la prueba global utilizada. Cuando hemos podido aplicar una prueba Anova de una vía, estándar, se sugiere el uso de una prueba *post hoc* HSD de Tukey. Si, por otro lado, fue necesario utilizar las correcciones de Brown-Forsythe o de Welch al Anova, deberá ser utilizada la prueba *post hoc* de Games-Howell. En tercer lugar, si utilizamos una prueba H de Kruskal-Wallis, la prueba *post hoc* adecuada es la prueba de Dunn. Finalmente, cuando sea necesario utilizar una prueba Chi-cuadrado, no hay una prueba *post hoc* específica, sino que debe procederse al análisis de residuos estandarizados.

El árbol de decisiones que describe todas las posibilidades que hemos mencionado aparece representado en la figura 58.

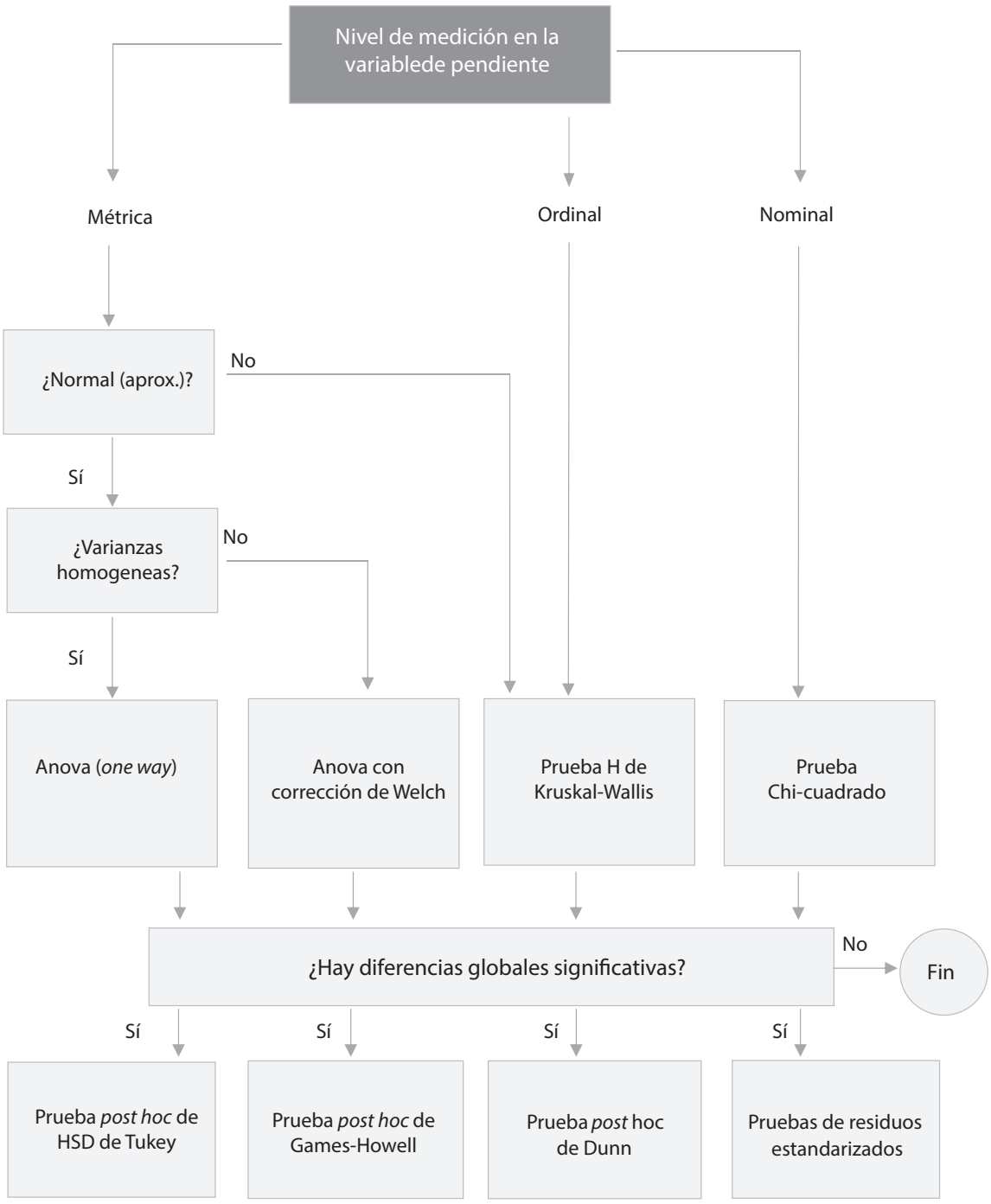


Figura 58. Diagrama de flujo para la selección de pruebas en k grupos independientes

Iniciamos con el caso en el que la variable dependiente es una variable métrica intervalar.

Variable métrica: el Anova en una dirección

Presentación

El *análisis de varianza en una dirección* (Anova *one way*) es la prueba paramétrica que permite examinar y comparar los niveles de significación de la diferencia entre *k* medias ($k \geq 2$). Si $k=2$, el Anova en una dirección es equivalente a una prueba *t* para grupos independientes.

La prueba requiere de una variable dependiente con nivel de medida métrico y continuo (por ejemplo, los resultados de una prueba) y una variable independiente que identifique los *k* grupos (por ejemplo, diferentes grupos escolares del mismo grado en un municipio).

El Anova ha sido descrito como una prueba de carácter general que compara si la varianza explicada es significativamente mayor que la varianza no explicada por la variable independiente. Proporciona un estadístico *F* (en honor a R. Fisher), dos valores correspondientes a los grados de libertad (gl1, o grados de libertad *entre* los grupos y gl2, o grados de libertad dentro de los grupos) y un nivel de significación asociado (*p*).

Los supuestos del Anova de una vía son los mismos de las demás pruebas paramétricas. A saber:

- La variable dependiente debe ser métrica y continua.
- La variable independiente debe describir grupos independientes entre sí (disyuntos)
- La variable dependiente debe tener una distribución aproximadamente normal y sin valores atípicos significativos.
- La variable dependiente debe mostrar homogeneidad de varianza entre los grupos. De otra forma, el estadístico y su valor de significación no son confiables.

Al respecto del supuesto de normalidad, debe anotarse que, aunque el Anova de una vía es una prueba robusta frente a la violación de este supuesto, si se requiere, al menos, que la distribución sea simétrica y no presente valores atípicos muy significativos. Cumplida esta condición, puede correrse la prueba. Si definitivamente esto no es posible, deberemos considerar hacer transformaciones sobre la variable o, en último caso, la prueba no paramétrica equivalente: la prueba H de Kruskal-Wallis, que será presentada más adelante.

Para el caso de la violación del supuesto de homogeneidad de varianzas, los diferentes paquetes estadísticos que utilizamos permiten el cálculo de dos correcciones al valor *F*, para usarlas en este caso: la *corrección de Brown-Forsythe* y la *corrección de Welch*. Específicamente, recomendamos el uso de la corrección de Welch para ser utilizada en el caso en el que se cumpla el supuesto de normalidad y se viole el supuesto de homogeneidad de varianzas. En esencia, estas dos correcciones modifican levemente los valores *F*, los grados de libertad (gl2) y los valores de la significación (*p*), si bien la corrección de Welch ha demostrado ser la más robusta (Frost, 2017).

La hipótesis nula que se pone a prueba en el Anova es que no hay diferencias entre las medias de los diferentes grupos. Si esta hipótesis se rechaza, el Anova afirmará que hay diferencias entre los grupos, pero no dirá cuáles son los grupos que marcan estas diferencias. Para conocerlos, así como sus diferencias con los otros, deben ser corridas pruebas *post hoc*.

Es muy importante que estas pruebas *post hoc* sean examinadas únicamente en el caso en el que el Anova de una vía haya mostrado ser significativo. De otra forma, tendríamos un aumento de la probabilidad del error de tipo I por acumulación de pruebas, por lo que los resultados no serían confiables. Ahora, para examinar las diferencias entre los grupos, existen muchas, y muy distintas pruebas *post hoc*, cuya selección dependerá, de nuevo, de si se cumplió o no el supuesto de homogeneidad de varianzas.

En el primer caso, en el que se cumple el supuesto de homogeneidad de varianzas, los diferentes paquetes ofrecen una gran cantidad de pruebas. En términos generales, se recomienda la selección de la *prueba de Tukey*, también conocida como la prueba HSD (*honest significant difference*) de Tukey o prueba honestamente significativa de Tukey. Otras opciones populares, viables y disponibles en todos los paquetes son las correcciones de Bonferroni y las de Scheffe**.** En JASP está disponible también una corrección de Holm, adicional a la de Boferroni, y en el SPSS aparecen catorce correcciones adicionales a las anteriores; un número exageradamente alto de pruebas para hacer la misma cosa.

Para el segundo caso, en el que no se cumpla el supuesto de homogeneidad de varianza, las opciones para pruebas *post hoc* son las pruebas T2 de Tamhane, T3 de Dunnet, Games-Howell y C de Dunnet en IBM-SPSS o las de Games-Howell y Dunn, en JASP. Para este caso se recomienda la *prueba de Games-Howell*, presente tanto en JASP como en IBM-SPSS. La prueba *post hoc* de Dunn, que está en JASP, es una prueba no paramétrica que se puede utilizar como prueba *post hoc* de la prueba no paramétrica *H* de Kruskal-Wallis, que expondremos más adelante.

En relación con las medidas de tamaño del efecto, el programa JASP permite tres opciones para el cálculo del tamaño del efecto en un Anova: el *eta cuadrado* (η^2), el *eta parcial al cuadrado* (η_p^2) y el *omega cuadrado* (ω^2)**.** El η^2 es una medida bastante popular, pero dificulta la comparación del efecto de la misma variable en distintos estudios, por lo que es preferible el uso del eta al cuadrado parcial (η_p^2) que resuelve este problema. Sin embargo, cuando las muestras son de tamaño pequeño ($n<30$), η_p^2 tiende a sesgarse, por lo que, en estos casos, se prefiere el uso del ω^2. La tabla 76 permite interpretar estas medidas de tamaño del efecto en el Anova de una vía.

Tabla 76. Límites para la interpretación de las medidas de tamaño del efecto en el Anova de una vía

Medida de tamaño del efecto	Nulo	Pequeño	Mediano	Grande
η^2	<0,1	0,1	0,25	0,37
η_p^2 (n>30)	<0,01	0,01	0,06	0,14
ω^2 (n<30)	<0,01	0,01	0,06	0,14

Fuente: Goss-Sampson (2019).

Por su parte, el IBM-SPSS no aporta el cálculo de ninguna medida de tamaño del efecto, si bien este podría ser calculado con una sencilla fórmula a partir de la información aportada en la tabla del Anova de una vía.

Los dos programas que utilizamos difieren levemente en la información que aportan. Ya mencionamos una ventaja importante en el JASP relacionada con la opción de calcular varios tamaños del efecto, que está ausente en el IBM-SPSS. Por otro lado, el IBM-SPSS tiene la ventaja de permitir examinar múltiples Anovas de una vía, de múltiples variables dependientes con la misma variable independiente en un solo procedimiento, lo que puede ser cómodo.

Ejecutar en Anova en una vía

Para examinar un análisis de varianza en una vía, en los diferentes programas, puede procederse de la siguiente forma: en el programa JASP, como se muestra en el recuadro 36; en el programa IBM-SPSSS, el procedimiento está presente en el menú "Comparar medias" (recuadro 37).

Recuadro 36. Ejecutar un Anova en una vía en JASP

/ANOVA/ANOVA.
En este punto debe pasarse la variable dependiente a la lista "dependent variable" y el factor independiente a la lista "fixed factors"
Display
- √ Descriptive statistics
- √ Estimates effect size
  - √ Partial η^2 o √ ω^2 (dependiendo del tamaño de la muestra)
- Assumption Checks
  - √ Homogeneity test
  - Homogenity corrections
    - (Seleccionar correcciones, dependiendo del test anterior)
    - √ None o √ Welch
  - √ Q-Q plot of residuals
- Post Hoc test
  - (dependiendo de la homogeneidad, se selecciona la prueba adecuada)
  - √ Tukey o √ Games-Howell
- Descriptive Plots (pasar la variable de factor a "Horizontal Axis")
  - √ Display error bars

Recuadro 37. Ejecutar un Anova en una vía en IBM-SPSS

/Analizar/Comparar medias/Anova de un factor…
En este punto se pueden pasar una o varias variables dependientes a la lista "Lista de dependientes" y elegir una variable independiente para trasladarla a "Factor"
En el botón "Post hoc"… se selecciona la prueba adecuada dependiendo de la prueba de homogeneidad de varianzas
√ Tukey o √ Games-Howell
Pulsar "Continuar"
En el botón "Opciones" es recomendable seleccionar:
√ Descriptivos
√ Prueba de homogeneidad de varianzas
√ Welch (solo se usará dependiendo de la prueba de varianzas)
√ Gráfico de medias
Pulsar "Continuar"
√ Estimar tamaño del efecto para pruebas generales
Pulsar "Aceptar"

El ejemplo: diferencias en el aprendizaje entre los niveles educativos

Un investigador se encuentra interesado en los cambios que viven los estudiantes a medida que avanzan en el sistema educativo. Para explorar la naturaleza de estos cambios aplica un instrumento que examina aspectos motivacionales y estratégicos en el aprendizaje a estudiantes de secundaria y de universidad en los niveles de pregrado y posgrado.

El instrumento aplicado es conocido como el cuestionario MSLQ (*motivated strategies for learning questionair*e) de Pintrich *et al.* (1993). Este cuestionario examina quince diferentes características de la motivación y la conducta durante el aprendizaje, puntuándolas en escalas numéricas continuas de 1 a 7. En la presente sección examinaremos las diferencias en tres escalas, a saber:

- *La autoeficacia académica*, relacionada con la confianza que tiene el estudiante en su propia capacidad académica.
- *La ansiedad evaluativa*, que indica la experiencia emocional del estudiante durante las evaluaciones.
- *La autorregulación metacognitiva*, que indica el grado de control del estudiante acerca de su propio proceso de aprendizaje.

En total, se cuenta con información de 151 estudiantes, de los cuales 79 (52,3 %) se encuentran cursando la secundaria, 50 (33,1 %) se encuentran en el pregrado universitario y 22 (14,6 %) adelantan estudios de posgrado. Interesa examinar si hay diferencias entre los niveles educativos en estas tres escalas.[8]

8 Tanto el proyecto de investigación como los datos concretos son reales y provienen de poblaciones de estudiantes de Bogotá, Colombia, y fueron previamente publicados, si bien los análisis estadísticos fueron modificados con propósitos pedagógicos. Los resultados originales pueden ser consultados en Hederich-Martínez *et al.* (2018).

Planteamiento de las hipótesis

Para este caso, examinaremos tres pruebas de Anova en una dirección, cada una para una escala: autoeficacia, ansiedad y autorregulación metacognitiva. Así las cosas, para la escala de autoeficacia, las hipótesis podrían ser formuladas de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en las medias de la escala de autoeficacia académica (ansiedad evaluativa/autorregulación metacognitiva) entre los diferentes niveles educativos.*
>
> *Hipótesis alternativa (H_1). Al menos uno de los niveles educativos tiene una media en la escala de autoeficacia académica (ansiedad evaluativa/autorregulación metacognitiva) significativamente diferente de los otros.*

Para la formulación de las hipótesis relativas a la escala de ansiedad evaluativa y a la autorregulación metacognitiva, se puede proceder de forma idéntica, cambiando el nombre de la escala, como aparece indicado en la redacción de las hipótesis.

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* Examinaremos el desarrollo de las tres pruebas de forma paralela. Esto significa que presentaremos, en cada paso, los resultados de las tres pruebas de Anova en una dirección para facilitar al lector la comparación de las diferentes pruebas. Esta no es la forma en que se presenta en el programa JASP.

El Anova en una vía requiere una variable dependiente numérica y continua, con una distribución aproximadamente normal y una variable independiente que defina dos o más grupos independientes. Para nuestro caso, la variable independiente es el nivel educativo, que define tres grupos independientes (secundaria, pregrado y posgrado). Por su parte, las variables dependientes están definidas por los puntajes en las escalas de autoeficacia, ansiedad evaluativa y autorregulación metacognitiva, que son variables numéricas continuas.

Para la verificación del supuesto de normalidad aproximada, en la figura 59, se presentan los gráficos Q-Q de las tres variables cuyas medias serán examinadas.

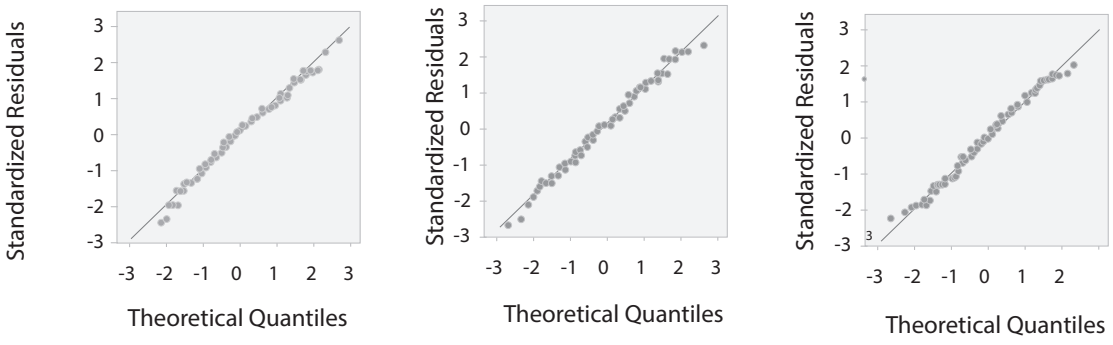


Figura 59. Gráficos Q-Q de autoeficacia, ansiedad y autorregulación metacognitiva

Como se observa en las tres ilustraciones, las líneas de puntos siguen muy de cerca la diagonal, con muy pequeñas desviaciones en uno o dos puntos al inicio o al final de la línea. Dado este resultado, podemos asumir que el supuesto de normalidad aproximada se cumple para las tres variables.

El segundo de los supuestos importantes para el análisis de varianza en una dirección es el supuesto de homogeneidad de varianzas. La tabla 77 muestra los resultados de la prueba de Levene para la verificación de este supuesto en cada una de las escalas.

Tabla 77. Resultados de las pruebas de Levene para la verificación del supuesto de igualdad de varianzas

Escala	F	gl1	gl2	p
Autoeficacia académica	1,493	2	148	,228
Ansiedad evaluativa	1,934	2	148	,148
Autorregulación metacognitiva	4,821	2	148	,009 **

De acuerdo con la tabla, los resultados de la prueba de Levene indican que el supuesto de homogeneidad de varianzas se cumple para el caso de las escalas de autoeficacia académica $F(2,148)=1,49$ $p=,228$ (ns) y ansiedad evaluativa $F(2,148)=1,93$ $p=,148$ (ns), pero no se cumple para el caso de la escala de autorregulación metacognitiva $F(2,148)=4,82$ $p=,009$. Esto indica que, mientras en los dos primeros casos podemos utilizar el Anova estándar, para el tercer caso debemos utilizar las correcciones de Brown-Forsythe o de Welch a esta prueba. Es importante tener este punto en cuenta en el momento de correr las pruebas.

• *Se examinan resultados descriptivos.* Las gráficas muestran las diferencias entre las medias de cada nivel educativo para cada una de las escalas que estamos examinando. Los resultados de los estadísticos descriptivos se presentan en la tabla 78 y deberán ser utilizados más adelante a la hora de expresar las medias de cada nivel educativo.

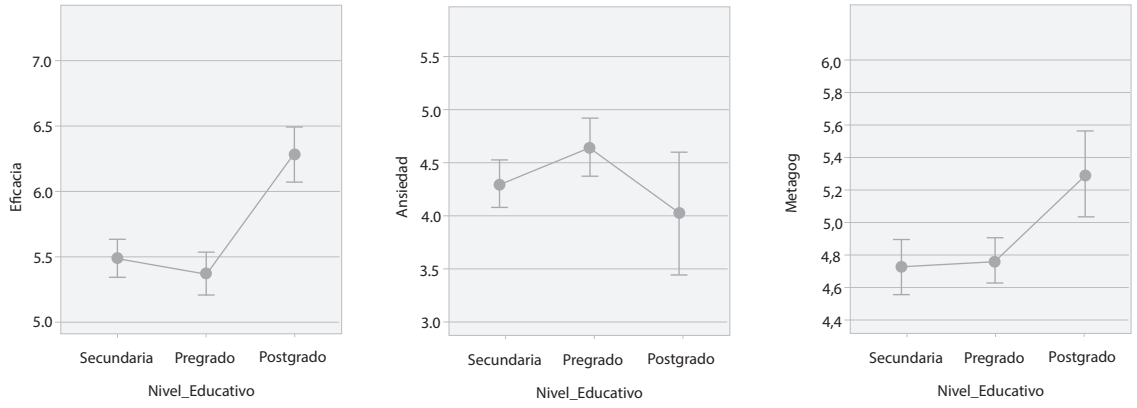


Figura 60. Medias y errores estándar de las diferentes escalas en los tres niveles educativos

Tabla 78. Descriptivos de las tres escalas en los tres niveles educativos

Nivel educativo	Autoeficacia			Ansiedad			Autorregulación		
	Media	DE	N	Media	DE	N	Media	DE	N
Posgrado	6,28	0,47	22	3,97	1,29	22	5,30	0,58	22
Pregrado	5,37	0,58	50	4,59	0,96	50	4,77	0,48	50
Secundaria	5,49	0,65	79	4,25	0,99	79	4,71	0;76	79

Como se observa, en la escala de autoeficacia académica los estudiantes de secundaria y pregrado muestran bajos puntajes, mientras que los de posgrado los manifiestan muy altos. Por su parte, la ansiedad evaluativa indica niveles medios en secundaria, un incremento en los estudiantes universitarios de pregrado y un brusco descenso en posgrado. Finalmente, para el caso de la autorregulación metacognitiva los resultados indican bajos puntajes en los estudiantes de secundaria y pregrado y un claro incremento para los estudiantes de posgrado. Los resultados de los análisis de varianza y, si proceden, de las pruebas *post hoc*, nos indicarán la significación de estas diferencias.

Se examinan los resultados de la prueba

La tabla 79 presenta el resultado del Anova en una vía para la escala de autoeficacia académica. Para este caso, y ya que tenemos una muestra mayor que 30 (n=151), solo se ha solicitado una, de las tres medidas disponibles, de tamaño del efecto: el eta cuadrado parcial (η_p^2).

Como se observa, el Anova muestra que las medias de la escala de autoeficacia difieren de forma significativa entre los niveles educativos (p<,001); el tamaño del efecto, por su parte, indica ser grande ($\eta_p^2 = 0{,}200 > 0{,}14$).

Tabla 79. Tabla del Anova de una vía para la escala de autoeficacia por nivel educativo

Anova–Autoeficacia						
Variable	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
Nivel_Educativo	1,683	2	6,841	18,466	< ,001	0,200
Residuos	54,832	148	0,370			

Note. Type III sum of squares.

Intentaremos hacer una breve explicación de cada uno de los elementos presentes en la tabla orientada a los interesados, aunque no es estrictamente necesaria para comprender el resultado del procedimiento.

Iniciando con la columna "Suma de cuadrados", en la primera fila (nivel educativo), esta presenta la suma de las diferencias entre la media de cada grupo (de nivel educativo) y la media general, al

cuadrado; esto es, cuánto varían los grupos. En la segunda fila representa la suma de las diferencias entre la medida de cada individuo y la media total, al cuadrado; esto es, cuánto es el resto de la variación de los individuos. La columna "gl" (df) representa la fuente de las variaciones en cada fila: el número de valores del nivel educativo menos uno en la primera, y el número de individuos menos tres en la segunda. La columna "Cuadrado medio" representa el cociente entre los anteriores dos valores para cada fila y, finalmente, la "F" representa el cociente entre los cuadrados medios.

Podemos añadir algo sobre la nota "Type III sum of squares". Existen cuatro tipos de sumas de cuadrados, rotulados I, II, III y IV. La suma de cuadrados tipo I se utiliza para diseños muy desequilibrados; el tipo II se utiliza cuando hay varios efectos principales; el tipo III es la opción por defecto, no depende del tamaño de la muestra; el tipo IV es una variación de la anterior que se utiliza en diseños con celdas faltantes.

Continuando ahora con el análisis, ya que sabemos que el análisis de varianza de la autoeficacia académica muestra diferencias significativas entre los niveles educativos, podemos examinar los grupos específicos que revelan diferencias a través de pruebas *post hoc*. Para este caso, ya que sabemos que esta variable evidencia homogeneidad de varianzas, entre los grupos, podemos examinar la prueba *post hoc* HSD de Tukey. Los resultados de estas pruebas se presentan en la tabla 80.

Tabla 80. Pruebas post hoc de Tukey para autoeficacia académica

Comparaciones *post hoc*–Nivel_Educativo							
			IC 95 % para la diferencia media				
		Diferencia media	**Bajo**	**Alto**	**EE**	**t**	**p_{tukey}**
Secundaria	Pregrado	0,118	-0,142	0,379	0,110	1,074	0,532
	Posgrado	-0,794	-1,141	-0,447	0,147	-5,412	< ,001 ***
Pregrado	Posgrado	-0,912	-1,281	-0,543	0,156	-5,857	< ,001 ***

*** p < ,001.

De acuerdo con los resultados, no hay diferencias significativas entre los estudiantes de secundaria y de pregrado en sus niveles de autoeficacia académica (p=,532) pero sí se encuentran diferencias significativas, a niveles inferiores a ,001, entre los de secundaria y posgrados, así como entre los de pregrado y posgrado. Los estudiantes de posgrado muestran mayores niveles de autoeficacia académica que los de secundaria y pregrado.

Continuamos ahora con el examen de los cambios en los niveles de ansiedad evaluativa. Como se recordará, la prueba de homogeneidad de varianzas nos indicó que esta escala cumple con este supuesto, lo que nos autoriza a examinar en Anova convencional sin correcciones adicionales. La tabla 81 muestra los resultados del análisis de varianza para esta escala.

Tabla 81. Tabla del Anova de una vía para la escala de ansiedad por nivel educativo

Variable	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
Nivel_Educativo	6,731	2	3,365	3,166	0,045	0,041
Residuos	157,338	148	1,063			

Como se observa, el análisis muestra diferencias significativas en la ansiedad evaluativa entre los niveles educativos, si bien los niveles de significación no resultan ser muy extremos (p=,045), apenas ,005 por debajo de los niveles convencionalmente aceptados. El tamaño del efecto, por su parte, indica ser más bien entre pequeño y mediano ($0{,}01 < \eta_p^2 = 0{,}041 < 0{,}06$). Véase la tabla 81.

Ya que el análisis de varianza de la ansiedad evaluativa mostró diferencias globales significativas, podemos proceder al examen de las pruebas *post hoc* entre los niveles educativos. El resultado de las comparaciones en las pruebas HSD de Tukey se presenta en la tabla 82.

Tabla 82. Pruebas post hoc para las diferencias en ansiedad evaluativa

Comparaciones *post hoc*–Nivel_educativo					
		Diferencia media	SE	t	p_{tukey}
Secundaria	Pregrado	-0,341	0,186	-1,832	0,163
	Postgrado	0,278	0,249	1,118	0,504
Pregrado	Postgrado	0,619	0,264	2,348	0,052

Los resultados son interesantes, por cuanto no parece haber una pareja de grupos de nivel educativo que muestre diferencias significativas. Los mayores contrastes se presentan entre los niveles de pregrado y posgrado, si bien su nivel de significación no alcanza el convencionalmente aceptado (p=,052>,050). Solo dos milésimas faltaron para el cumplimiento de este requisito.

Esta situación amerita que se recuerde el carácter convencional del nivel de significación elegido. ¿Dado que no alcanzamos, por dos milésimas, al nivel de ,05, entonces, debemos concluir que no hay diferencias significativas? No creo. Por un lado, el Anova ya había mostrado diferencias globales significativas. Por el otro, la proximidad al nivel de significación convencionalmente aceptado nos indica la presencia de diferencias importantes entre estos dos grupos. Desde este punto de vista, puede aceptarse la presencia de diferencias entre los niveles de pregrado y posgrado en relación con la ansiedad evaluativa experimentada.

Por último, debemos examinar las diferencias entre las medias de la escala de autorregulación metacognitiva entre los tres niveles educativos. Como se recordará, la prueba de Levene de homogeneidad de varianzas rechazó la hipótesis nula para esta escala, por lo que no pudimos verificar este supuesto. Por esta razón, debemos examinar el análisis de varianza con alguna de las correcciones ofrecidas por el *software* para este caso: las correcciones de Brown-Forsythe y de Welch.

La tabla 83 muestra los resultados del análisis de varianza con estas dos correcciones, además del Anova sin correcciones, que se incluye aquí con propósitos pedagógicos.

Tabla 83. Tabla del Anova convencional, y con corrección, para la autorregulación metacognitiva

Anova – Autorregulación metacognitiva							
Corrección de Homogeneidad	**Variables**	**Suma de cuadrados**	**gl**	**Cuadrado medio**	**F**	**p**	**η^2_p**
Ninguna	Nivel_Educativo	6,21	2,00	3,10	7,25	< ,001	0,089
	Residuos	63,38	148,00	0,43			
Brown-Forsythe	Nivel_Educativo	6,21	2,00	3,11	8,59	< ,001	0,089
	Residuos	63,38	94,26	0,67			
Welch	Nivel_Educativo	6,21	2,00	3,11	8,45	< ,001	0,089
	Residuos	63,3	59,80	1,06			

Nota: suma de cuadros tipo III.

De acuerdo con la tabla, los resultados indican diferencias globales significativas entre los niveles educativos en la escala de autorregulación metacognitiva (p<,001 en cualquiera de las correcciones al Anova, así como en el Anova sin corrección). Cualquiera de los tres análisis conduce a la misma conclusión, aunque formalmente solo deberíamos reportar alguna de las correcciones. Los tamaños del efecto, por su parte, son entre medianos y grandes ($0,06 < \eta_p^2 = 0,089 < 0,14$). Es interesante observar que las medidas de tamaño del efecto muestran ser iguales para todas las pruebas.

El hecho de que, en este caso, el Anova muestre diferencias globales significativas nos autoriza a examinar las pruebas *post hoc*, a fin de descubrir los grupos específicos que evidencian diferencias entre sí. Para hacerlo, debemos recordar, de nuevo, que en esta escala la prueba de Levene señaló que no se puede sostener el supuesto de homogeneidad de varianzas lo que nos indica que debemos correr pruebas *post hoc* no estándares; específicamente, examinaremos los resultados de las comparaciones *post hoc* de Games-Howell, que resultan apropiadas para este tipo de casos. Los resultados se presentan en la tabla 84.

Tabla 84. Comparaciones post hoc de Games-Howell para autorregulación metacognitiva

Comparaciones *post hoc* de Games Howell–Nivel_Educativo					
Comparación	**Diferencia media**	**EE**	**t**	**gl**	**p_{tukey}**
Secundaria–Pregrado	-0,061	0,109	-0,561	126,970	,841
Secundaria–Posgrado	-0,593	0,151	-3,926	42,584	<,001 ***
Pregrado–Posgrado	-0,532	0,142	-3,735	34,234	,002 **

** p < .01, *** p < .001.

Los resultados muestran diferencias significativas a favor de los estudiantes de posgrado frente a los de secundaria y pregrado (p<,001 y p=,002, respectivamente), mientras que no se presentan diferencias significativas entre los estudiantes de secundaria y los de pregrado en esta escala (p=,841).

Se expresan los resultados

Por la gran cantidad de resultados, puede ser recomendable incluir las gráficas contenidas en la figura 59, o los descriptivos de la tabla 76, para presentar los resultados de medias y desviaciones estándar de los diferentes niveles educativos. Hecho esto, las pruebas que examinan las diferencias entre los niveles pueden ser expresadas en texto.

Para expresar los resultados del Anova en una vía, en texto, puede usarse el siguiente formato:

F(<gl1>,<gl2>)= <valor F> p =<Valor p> $\eta^2/\eta_p^2/\omega^2$ = <valor tamaño del efecto>

Utilizando este formato, los resultados previos pueden ser expresados, en texto, de la siguiente forma:

> *Se utilizó la prueba Anova en una vía para examinar las diferencias en las medias de las escalas de autoeficacia académica, ansiedad evaluativa autorregulación metacognitiva entre los estudiantes de secundaria, pregrado y posgrado. El análisis de los supuestos de homogeneidad de varianzas a través de la prueba de Levene indicó que las varianzas pueden ser consideradas homogéneas para las escalas de autoeficacia académica F(2,148)=1,49 p=,228 y ansiedad evaluativa F(2,148)=1,93 p=,148, pero esto no puede ser aseverado en la escala de autorregulación metacognitiva F(2,148)=14,82 p=,009, por lo que se requiere, en este último caso, el uso de pruebas robustas de Welch con comparaciones post hoc de Games-Howell, para corregir el efecto de la falta de homogeneidad de varianzas.*
>
> *Los resultados de los Anova indicaron diferencias significativas entre los niveles educativos para las escalas de autoeficacia académica, con tamaños de efecto grandes F(2,148)=18,47 p<,001 η_p^2=0,20; ansiedad evaluativa, con tamaños de efecto entre pequeños y medianos F(2,148)=3,17 p=,045 η_p^2=0,04 y autorregulación metacognitiva con tamaños del efecto entre medianos y grandes F(2, 59,81)=8,45 p<,001 η_p^2=0,09.*
>
> *Las pruebas post hoc HSD de Tukey indicaron que el nivel de posgrado muestra una autoeficacia académica significativamente más alta que los de secundaria (p<,001) y pregrado (p<,001), mientras que no se encuentran diferencias entre secundaria y pregrado (p=,532). Por su parte, los mayores niveles de ansiedad evaluativa se hallan en estudiantes de pregrado, que alcanzan a manifestar diferencias apreciables no significativas con los del posgrado (p=,052); no se encuentran diferencias significativas entre secundaria y pregrado (p=,163), ni entre secundaria y posgrado (p=,504). Al respecto de la autorregulación metacognitiva, las pruebas* post hoc *de Games-Howell revelaron que el nivel de posgrado muestra puntajes significativamente más altos que los de pregrado (p=,002) y secundaria (p<,001), mientras que no se encuentran diferencias significativas entre estos últimos dos niveles (p=,841).*

Variable ordinal: prueba H de Kruskal-Wallis

Presentación

Cuando queremos examinar las diferencias entre tres grupos o más y la variable dependiente tiene un nivel de medida ordinal, o muestra una seria violación del supuesto de normalidad, la prueba adecuada es la prueba H de Kruskal-Wallis.

La *prueba H de Kruskal-Wallis* es la alternativa no paramétrica al Anova de una vía, de la misma forma en que la prueba U de Mann-Whitney era la alternativa no paramétrica a la prueba *t* de Student; de hecho, la prueba H de Kruskal-Wallis se puede considerar como una extensión de la prueba U de Mann-Whitney para tres o más grupos.

La prueba H de Kruskal-Wallis es una prueba de rangos, y por tanto, compara los rangos promedio de los diferentes grupos de la muestra. Produce un estadístico (*H*) y un valor de la significación asociado. La hipótesis nula de esta prueba plantea que todas las muestras provienen de la misma población y por tanto tienen la misma distribución, mientras que la hipótesis alternativa plantea que al menos una muestra proviene de una población con una distribución distinta.

Como esta es una prueba no paramétrica, no requiere de los supuestos usuales de las paramétricas en lo relacionado con la normalidad. Sin embargo, en el caso en el que tengamos una variable métrica, una muestra muy pequeña y no se cumpla con la homogeneidad de varianzas, sería preferible recurrir al Anova con corrección de Welch que presentamos en la sección anterior. Salvo este caso particular, el único requisito para la prueba es una variable dependiente con nivel de medida, como mínimo, ordinal, y una variable independiente que defina dos o más grupos independientes (disyuntos).

Al igual que el Anova en una dirección, la prueba H de Kruskal-Wallis informa acerca de diferencias globales entre los grupos, pero no indica los grupos específicos que presentan diferencias significativas entre sí. Para saberlo es necesario, como en el caso del Anova, correr pruebas *post hoc*. Debe subrayarse que este tipo de pruebas solo deben ser examinadas en el caso en el que la prueba general señale diferencias globales significativas.

Las pruebas *post hoc* disponibles difieren entre los programas estadísticos que manejamos. En el caso del JASP, se propone la *prueba de Dunn*. Esta es una prueba no paramétrica perfectamente adecuada para el análisis *post hoc* de diferencias entre cada pareja de grupos en variables ordinales, que compensa los problemas derivados de establecer múltiples comparaciones. Cuando se solicita esta prueba, el programa aporta también los niveles de significación con corrección de Bonferroni y de Holm, que resultan más conservadores que los de la misma prueba de Dunn.

Para el caso del IBM-SPSS, el procedimiento de la prueba de Kruskal-Wallis no da ninguna posibilidad de seleccionar una prueba *post hoc* apropiada. El manual del programa sugiere, para solucionar este interrogante, examinar pruebas U de Mann-Whitney entre los diferentes grupos, pero sabemos que este procedimiento tiene el problema de aumentar el error de tipo I por acumulación de pruebas, por lo que habría que extremar los niveles de significación aceptados. Es una lástima que el IBM-SPSS no ofrezca esa posibilidad.

Respecto de las medidas de tamaño del efecto, debe anotarse que no hay una medida ni un procedimiento consensualmente aceptado para expresar el tamaño del efecto en una prueba H de

Kruskal-Wallis. Tratando de solucionar esta situación, Tomczak y Tomczak (2014) proponen dos posibles medidas de tamaño del efecto: el h^2_H y el ε^2_R. A partir de sus propuestas nos permitimos sugerir el uso del estadístico épsilon al cuadrado (ε^2_R), como una buena medida para la estimación del tamaño del efecto en esta prueba.[9] El ε^2_R se define de la siguiente forma:

$$\varepsilon^2_R = \frac{H}{(n-1)}$$

Donde H es el estadístico de Kruskal-Wallis y *n* es el número total de casos en el estudio.[10]

Para la interpretación de este estadístico pueden seguirse las recomendaciones de Cohen para la interpretación del valor *r* como medida de tamaño del efecto.

Ejecutar la prueba H de Kruskal-Wallis

Los dos programas que utilizamos difieren de forma importante en la información que aportan al correr una prueba de Kruskal-Wallis. Ya mencionamos una ventaja importante en el JASP relacionada con la presencia de la prueba de Dunn como prueba *post hoc* apropiada. En el sentido negativo, la prueba Kruskal-Wallis aparece como opción "no paramétrica" en el menú del Anova, junto con una gran cantidad de opciones que no resultan adecuadas para esta prueba (descriptivas apropiadas solo para variables métricas, tamaños del efecto de Anova, etc.). No se entrega en este programa información de rangos, que sería más apropiada para variables ordinales.

Por el otro lado, el uso de IBM-SPSS tiene la ventaja de aportar tablas sobre rangos, muy apropiadas para verificar diferencias entre variables ordinales. De igual forma, en este programa es posible examinar varias pruebas con varias variables dependientes y una misma variable independiente en un solo procedimiento lo cual puede ser relativamente cómodo. Como desventaja, que ya mencionamos, está la ausencia de pruebas *post hoc* apropiadas.

Tanto en el JASP como en el IBM-SPSS se carece de una medida adecuada para la cuantificación del tamaño del efecto, pero esto puede ser fácilmente subsanado a partir del uso del ε^2_R que hemos sugerido.

Para correr la prueba H de Kruskal-Wallis, en los diferentes programas, puede procederse de la siguiente forma. En el programa JASP, como se muestra en el recuadro 38; en el programa IBM SPSS, según el recuadro 39.

9 Las propuestas de Tomczak y Tomczak (2014) fueron puestas a prueba, a través de la ejecución de diferentes simulaciones, por el profesor C. Lanziano (comunicación personal). Sus resultados señalan la medida del *ε2* como la más adecuada para la indicación del tamaño del efecto en una prueba H de Kruskal-Wallis en tanto resulta más estable en la estandarización entre los valores 0 y 1.

10 Esta fórmula se presenta simplificada respecto de la original, que aparecía innecesariamente compleja.

Recuadro 38. Cómo ejecutar una prueba H de Kruskal-Wallis en JASP

/ANOVA/ANOVA..
En este punto debe pasarse la variable dependiente a la lista "dependient variable" la variable independiente a la lista "fixed factors"
Display
√ Descriptive statistics
Post Hoc test
√ Dunn
Nonparametrics
(pasar la variable de factor al cuadro correspondiente)

Recuadro 39. Cómo ejecutar una prueba H de Kruskal-Wallis en IBM-SPSS

/Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/K muestras independientes…
En este punto se pueden pasar varias variables dependientes a la lista "Lista variables de prueba" y elegir una variable independiente para todas las dependientes. El rango de la variable independiente debe ser especificado.
√ H de Kruskal-Wallis.
Pulsar "Aceptar"

El ejemplo: diferencias en el clima escolar entre las jornadas educativas

En un estudio existe el interés de examinar diferencias entre las jornadas escolares (de la mañana, de la tarde o jornada única) al respecto del clima escolar. El clima escolar es una variable construida con multitud de indicadores físicos y de relaciones entre docentes y estudiantes dentro de la institución, que finalmente se expresa con un nivel de medida ordinal de cinco puntos (muy negativo, negativo, neutro, positivo, muy positivo).

A fin de ilustrar el uso de la prueba de Kruskal-Wallis en situaciones en donde no se rechaza la hipótesis nula, se examinará el contraste entre las diferentes jornadas escolares respecto de la funcionalidad de las familias de los estudiantes que a ellas asisten, en cuanto al apoyo brindado. El cuestionario de funcionalidad familiar es una prueba de autopercepción del bienestar y del apoyo familiar que, después de combinar múltiples indicadores, se expresa como una variable ordinal de cuatro puntos (severa, moderada, leve, buena). Para esa variable no se espera que haya diferencias entre las jornadas.

Así, se dispone de información de 3568 estudiantes de diferentes instituciones educativas públicas en Colombia. Estos datos provienen de un estudio real, si bien su procesamiento fue adaptado con propósitos de ilustración de la prueba.

Planteamiento de las hipótesis

Para este caso, examinaremos dos grupos de hipótesis, el primero relacionado con el clima escolar y el segundo con el cuestionario de funcionalidad familiar. Así las cosas, las hipótesis sobre las diferencias en los niveles de clima escolar podría ser como sigue:

> *Hipótesis nula (H_0). No hay diferencias en las distribuciones del nivel de clima escolar entre las diferentes jornadas educativos (mañana, tarde y única).*
>
> *Hipótesis alternativa. H_1. Al menos una de las jornadas educativas (mañana, tarde o única) tiene una distribución del nivel de clima escolar diferente de las otras.*

Al formular estas hipótesis en términos de las diferencias entre los niveles de funcionalidad familiar, o apoyo familiar percibido, quedarían de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en las distribuciones del nivel funcionalidad familiar entre las diferentes jornadas educativas (mañana, tarde y única).*
>
> *Hipótesis alternativa. (H_1). Al menos una de las jornadas educativas (mañana, tarde o única) tiene una distribución del nivel de funcionalidad familiar diferente de las otras.*

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* La prueba H de Kruskal-Wallis solo requiere una variable dependiente con un nivel de medida, al menos, ordinal, y una variable independiente que defina grupos disyuntos. Tanto los niveles de clima escolar como los de funcionalidad familiar cumplen el supuesto de tener un nivel de medida ordinal, mientras que la jornada escolar cumple con el supuesto de ser una variable que define factores disyuntos. Podemos aplicar la prueba H de Kruskal-Wallis para el examen de las diferencias entre las jornadas escolares.
- *Se examinan resultados descriptivos.* La tabla 85 presenta el cruce entre el clima escolar percibido y la jornada escolar. La gráfica de la figura 61 presenta esta misma información como gráfica de barras apiladas al 100 %.

Tabla 85. Tabla de cruce entre clima escolar y jornada

Clima escolar	Jornada			Total
	Mañana	Tarde	Única	
Muy negativo	403	197	130	730
Negativo	363	219	123	705
Neutro	385	224	114	723
Positivo	348	237	109	694
Muy positivo	367	244	105	716
Total	1866	1121	581	3568

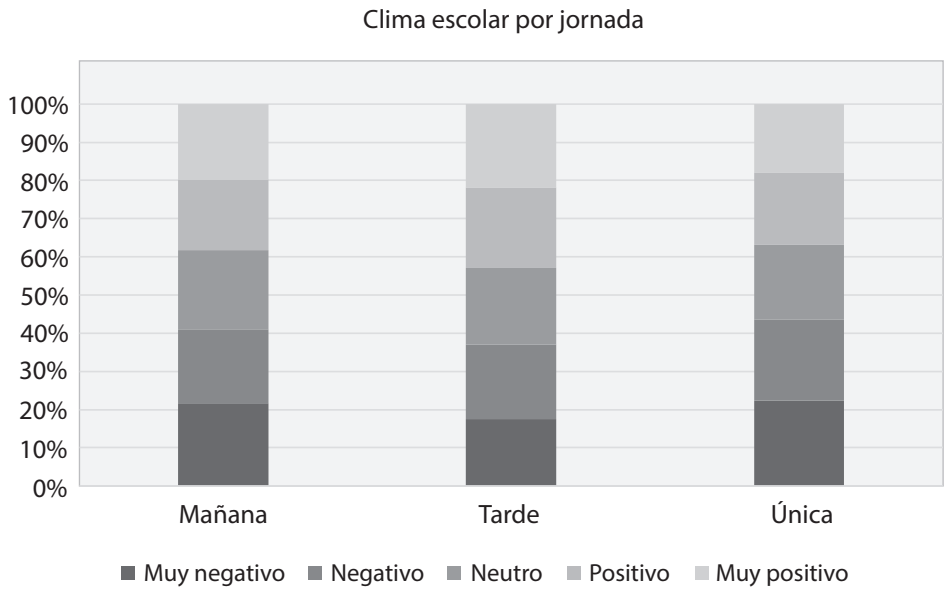


Figura 61. Gráfica de barras apiladas al 100 % del cruce entre clima escolar y jornada

Una inspección visual de la gráfica muestra diferencias relativamente leves entre las jornadas, que indican que la jornada de la tarde contiene una mayor proporción de estudiantes que presentan una percepción muy positiva del clima escolar y, al tiempo, una menor proporción de estudiantes que sienten un clima escolar muy negativo, en comparación con las otras dos jornadas escolares. Las diferencias, sin embargo, no parecen ser demasiado pronunciadas. Solo la prueba podrá indicarnos que tan significativa puede ser esta diferencia.

Al respecto de la segunda de las variables dependientes que examinaremos, la tabla 86 y la gráfica de la figura 62 muestran el cruce entre la funcionalidad familiar (apoyo familiar percibido por los estudiantes) y la jornada escolar.

Tabla 86. Tabla de cruce entre funcionalidad familiar y jornada

Funcionalidad familiar	Jornada			Total
	Mañana	Tarde	Única	
Severa	228	142	73	443
Moderada	450	251	142	843
Leve	777	451	228	1456
Buena	411	277	138	826
Total	1866	1121	581	3568

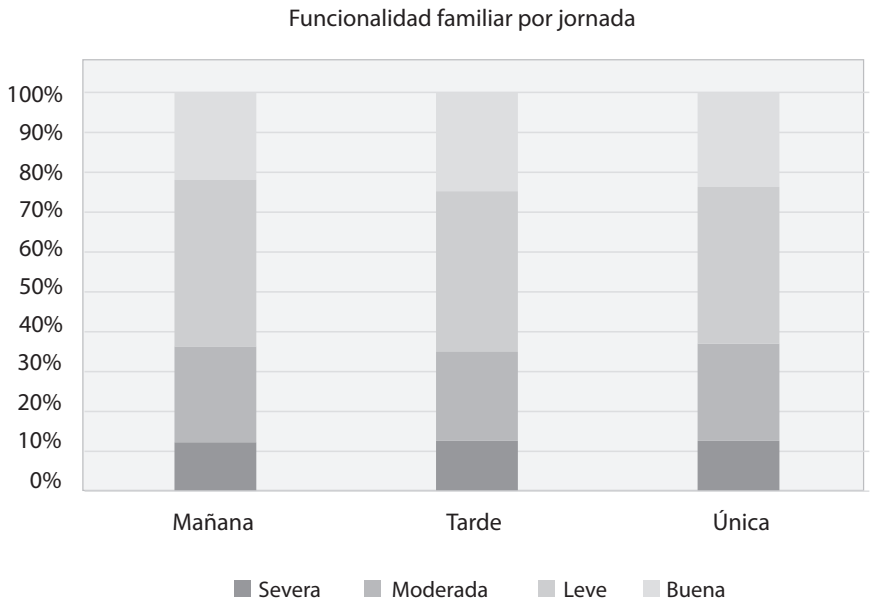


Figura 62. Gráfica be barras apiladas al 100 % del cruce entre funcionalidad familiar y jornada

Para este caso, las diferencias parecen ser mucho más sutiles. Si aguzamos la mirada, podría notarse que la categoría "buena" en la jornada de la tarde parece ser levemente más amplia que la misma en la jornada de la mañana. Un examen más detallado implicaría analizar los porcentajes para cada columna. Veremos lo que muestran los resultados de la prueba.

Se examinan los resultados de la prueba

Iniciaremos con el examen de los resultados de la prueba del clima escolar contra la jornada. La tabla 87 muestra que existen diferencias significativas en los niveles de clima escolar entre las jornadas (p=,004).

Tabla 87. Resultado de la prueba de Kruskal-Wallis de clima escolar por jornada

Factor	Estadístico	gl	p
Jornada	10,88	2	,004

Calculado el estadístico H de la prueba, podemos calcular el tamaño del efecto utilizando para ello la fórmula del ε^2_R:

$$\mathcal{E}^2_R = \frac{H}{(n-1)} = \frac{10{,}876}{3567} = 0{,}0030$$

El resultado general de la prueba nos autorizó al examen de las pruebas *post hoc*. La tabla 88 muestra los resultados de las pruebas *post hoc* de Dunn para el examen de las diferencias entre cada pareja de jornadas. Tal y como se observa, existen diferencias significativas entre la jornada de la

mañana y la de la tarde (p=,003), así como entre la jornada de la tarde y la jornada única (p=,002), mientras que no se verifican diferencias significativas entre la jornada de la mañana y la jornada única (p=0,173).

Tabla 88. Pruebas post hoc de Dunn para la comparación del clima escolar entre las jornadas

Comparación	z	W_i	W_j	p
Mañana–Tarde	-2,73	1759,07	1863,40	,003 **
Mañana–Única	0,94	1759,07	1713,91	,173
Tarde–Única	2,89	1863,40	1713,91	,002 **

** p < ,01.

Para comprender el sentido de las diferencias, puede notarse que la comparación entre jornada de la mañana y de la tarde muestra un valor z de -2,73, negativo, lo que indica que la primera de ellas —la de la mañana— presenta un nivel menor, o inferior, a la segunda, de la tarde; esto, en el presente contexto, significa que la jornada de la mañana tiene un clima menos positivo que la de la tarde. Otra forma de verificarlo es a través del examen de los W_i y W_j. El primero de ellos representa el rango promedio de la jornada de la mañana (1759,07) que resulta menor que el segundo, que presenta el rango promedio de la jornada de la tarde (1863,40).

Por su parte, la comparación entre la jornada de la tarde y la única muestra un valor z positivo, que ahora indica que la jornada de la tarde presenta un clima más positivo que la jornada única. En conclusión, la jornada de la tarde tiene un clima escolar significativamente más positivo que las jornadas de la mañana y única.

En la tabla 88 se presentan los niveles de significación con las correcciones de Bonferroni y Holm. Como se observa, en los dos casos estos niveles son levemente mayores a los obtenidos en la prueba original, pero conducen en todos los casos a las mismas conclusiones. Se evidencia que estas dos correcciones resultan ser un poco más conservadoras que la prueba original de Dunn. Finalmente, debemos examinar las diferencias entre los niveles de funcionalidad familiar percibida y las jornadas escolares. Al respecto, los resultados de la prueba H de Kruskal-Wallis se presentan en la tabla 89.

Tabla 89. Resultado de la prueba de Kruskal-Wallis de funcionalidad familiar por jornada

Factor	Estadistico	gl	p
P3_Jornada	1,332	2	,514

Como se observa, los resultados globales de la prueba no muestran diferencias significativas entre las distribuciones de la funcionalidad familiar por jornada escolar, por lo que debe aceptarse la hipótesis nula. Este resultado no nos autoriza al examen de pruebas *post hoc.*

Se expresan los resultados

Para la expresión de los resultados de la prueba H de Kruskal-Wallis puede seguirse este formato:

H(<gl>) = <valor z> *p* </= <valor p>

Los resultados de las pruebas *post hoc* pueden ser expresados solo en términos del nivel de significación alcanzado.

Siguiendo estas convenciones, los resultados de las dos pruebas pueden ser expresados en texto de la siguiente forma:

> *Los resultados de la aplicación de la prueba H de Kruskal-Wallis al cuestionario de clima escolar muestra diferencias significativas entre las jornadas escolares H(2)=10,88 p=,004. Las comparaciones post hoc indican que la jornada de la tarde tiene niveles de clima escolar más positivos que los de la mañana (p=,003) y que los de la jornada única (p=,002). No hay diferencias significativas entre la jornada de la mañana y la jornada única (p=,173).*
>
> *Por su parte, la prueba no señaló diferencias significativas entre las jornadas escolares en relación con el nivel de funcionalidad familiar reportado por los estudiantes H(2)= 1,33 p=,514.*

Variable nominal: Chi-cuadrado (χ^2) de Pearson en tablas de contingencia

Presentación

Presentaremos ahora la situación de la comparación entre *k* grupos respecto de una variable de naturaleza estrictamente nominal.

Tal y como lo hicimos para el caso de la comparación de dos grupos, que examinamos en el capítulo 8, para el examen de las diferencias entre k grupos (k>2) utilizaremos la misma prueba: *Chi-cuadrado (χ^2) de Pearson* en tablas de contingencia. La única diferencia entre la situación en la que comparábamos dos grupos y la presente es que ahora la variable independiente, que define la pertenencia a los grupos, no tiene ya dos valores sino "k" valores.

De resto, en todos los aspectos y los criterios de decisión nos mantenemos iguales. Recordaremos aquí los puntos principales.

La hipótesis nula de la prueba χ^2 de Pearson la de la independencia de las dos variables. Por el contrario, la hipótesis alternativa dirá que hay una relación de dependencia o asociación entre las dos variables.

Existen restricciones al uso de la prueba χ^2 de Pearson que es importante recordar:

- La prueba no debe ser corrida en muestras pequeñas (n<30).
- La prueba no debe ser corrida si existen demasiadas casillas, en las tablas de cruce, con una frecuencia menor que cinco. Como máximo, se admitiría que el 20 % de las celdas tengan una frecuencia menor que cinco, pero ninguna puede tener una frecuencia menor que uno.

En el caso en que se incumpla alguna de las restricciones, existen varias alternativas. En muestras pequeñas, se puede usar la "corrección por continuidad", la prueba exacta de Fisher, si está disponible o, mejor aún, la razón de verosimilitud (*likelihood ratio*).

Respecto de las medidas de tamaño del efecto, recomendamos el uso de la *V* de Cramer, cuyo valor oscila entre 0 y 1. Copiamos acá la tabla para la interpretación de la *V* de Cramer, dependiendo de la dimensión de la tabla.

Tabla 90. Interpretación de los valores de ϕ y V dependiendo de gl

Tamaño del efecto	gl* (df)	Pequeño	Mediano	Grande
ϕ y *V* de Cramer	1	0,10	0,30	0,50
V de Cramer	2	0,07	0,21	0,35
	3	0,06	0,17	0,29
	4	0,05	0,15	0,25
	5	0,04	0,13	0,22

* Los grados de libertad (gl) dependen de la dimensión de la tabla. En una tabla de dimensión r*c, los grados de libertad serán gl= (r-1)*(c-1)

Fuente: Kim (2017) y Goss-Sampson (2019).

Un punto más. Para el caso que nos ocupa, puede ser difícil, en ocasiones, estimar la magnitud de los cambios entre los diferentes valores de la variable independiente (esto es, entre los diferentes grupos que se quieren contrastar). Para hacerlo, el procedimiento clásico supone el cálculo y la valoración de los *residuos estandarizados* en cada una de las casillas de la tabla de cruce.

Para la valoración de las diferencias entre los grupos el criterio es simple: si en una casilla el residuo estandarizado es mayor, en valor absoluto, a 1,96, diremos que hay una diferencia significativa en ese valor de ese grupo específico. La comparación con 1,96 se da por cuanto este valor representa la frontera, en la distribución normal, para el logro de un nivel de significación de p=,05. El signo del residuo estandarizado nos dirá si ese valor cambia entre los grupos en el sentido en que se pierden casos (negativo) o, por el contrario, en el que se ganan casos (positivo).

No todos los programas utilizados calculan y presentan los residuos estandarizados. Específicamente, el programa JASP no lo hace, mientras que el programa IBM-SPSS lo hace de dos formas: como "residuos estandarizados" o, en las nuevas versiones, como "residuos estandarizados corregidos". En los dos casos es prácticamente equivalente.

Ejecutar la prueba χ2 de Pearson

Para el cálculo de estas pruebas, puede procederse, de la misma forma que lo hicimos al comparar dos grupos con una variable dependiente nominal. En JASP, esto se hace de la siguiente forma:

Recuadro 40. Cómo ejecutar la prueba χ^2 de Pearson en JASP

/Frequency/Contingency Tables
En este punto, deben seleccionarse la variable dependiente (y pasarse a la lista "Filas" y la variable independiente, pasándola a la casilla "Columnas" (en esta sección debería ser una variable con solo dos valores pero admitirá cualquier variable categórica).
Statistics (debe seleccionarse la prueba adecuada al tamaño de muestra).
√ χ^2
√ χ^2 continuity correction
√ Likelihood ratio
Nominal
√ Phi and Cramer's V

En el IBM-SPSS existe una posibilidad no ofrecida por el JASP que consiste en un examen *post hoc* a partir del cálculo de los residuos estandarizados corregidos. Las especificaciones son como sigue:

Recuadro 41. Cómo ejecutar la prueba χ^2 de Pearson en IBM-SPSS

/Analizar/Estadísticos descriptivos/Tablas cruzadas…
En este punto, deben seleccionarse la variable dependiente (y pasarse a la lista "Filas" y la variable independiente, pasándola a la casilla "Columnas" (para esta sección debería ser una variable con solo dos valores pero admitirá cualquier variable categórica).
En el botón "Estadísticos"
√ Chi-cuadrado
Nominal
√Phi y V de Cramer
Pulsar el botón "Continuar"
En el botón "Casillas"
Residuos
√ Estandarizados, o √ Estandarizados corregidos
Pulsar el botón "Continuar"
Pulsar el botón "Aceptar"

El ejemplo: diferencias en la actividad económica de los egresados de cuatro programas

Diferentes instituciones han diseñado e implementado una serie de programas educativos que pretenden facilitar a los jóvenes que previamente han desertado del sistema educativo la culminación de sus estudios al nivel de secundaria. Dentro de los muchos criterios de comparación posibles, se examinará la actividad desarrollada por los estudiantes egresados de cada uno de los programas un año después de su grado, en términos de una variable nominal, relacionada con la actividad económica del egresado, a saber:

Actividad desarrollada por el estudiante, un año después de su grado:

- Estudio
- Trabajo remunerado
- Estudio y trabajo
- Ni estudian ni trabajan (Ni-Ni).

Para este ejemplo, compararemos los resultados de los supuestos programas educativos de cuatro instituciones, rotuladas como "Institución 1", "Institución 2", "Institución 3" e "Institución 4".

Contamos con información de 348 egresados de cuatro instituciones (datos ficticios).

Planteamiento de las hipótesis

En el ejemplo que se desarrolla, los investigadores formularían las hipótesis de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en la actividad económica a un año de terminado el programa entre los egresados de los cuatro programas educativos examinados.*
>
> *Hipótesis alternativa (H_1). Existen diferencias en la actividad económica a un año de terminado el programa entre los egresados de los cuatro programas educativos examinados.*

Se corre la prueba

Se examinan los supuestos y se selecciona la prueba. La prueba c^2 de Pearson requiere de dos variables con nivel de medida categorial, cada una con un mínimo de dos valores, que sean temporalmente independientes; esto es, que no correspondan a un diseño intrasujeto, del tipo antes y después.

Por otro lado, la prueba requiere de un tamaño mínimo de muestra mayor que treinta casos, sin celdas nulas en el cruce de variables, y con una frecuencia de más de cinco casos en, al menos, el 80 % de las celdas.

Para nuestro caso, contamos con dos variables temporalmente independientes, las dos con cuatro valores diferentes. El tamaño de muestra es de 348 casos; ninguna celda en el cruce de variables presenta menos de siete casos (ver tabla 90), por lo que los supuestos de distribución de la muestra se cumplen.

- Se examinan resultados descriptivos. La tabla 91 es la de cruce entre las dos variables.

Para la valoración de las diferencias puede hacerse la inspección visual de la gráfica de barras apiladas al 100 %, presentada en la figura 63. Como se observa a simple vista, la institución 1 parece mostrar más egresados que continúan estudiando y menos que estudian y trabajan, y la institución 3 parece presentar más egresados que estudian y trabajan, mientras que la institución 4 parece tener mayor proporción de estudiantes Ni-Ni, y menor proporción de estudiantes que trabajan.

Tabla 91. Tabla de contingencia entre la actividad económica, a un año del grado, y el programa seguido (institución)

		Actividad (1 año)				Total
		Estudia	Trabaja	Est y Tra	Ni-Ni	
Institución	Inst 1	29	31	18	16	94
	Inst 2	22	30	28	19	99
	Inst 3	9	27	41	16	93
	Inst 4	14	7	19	22	62
Total		74	95	106	73	348

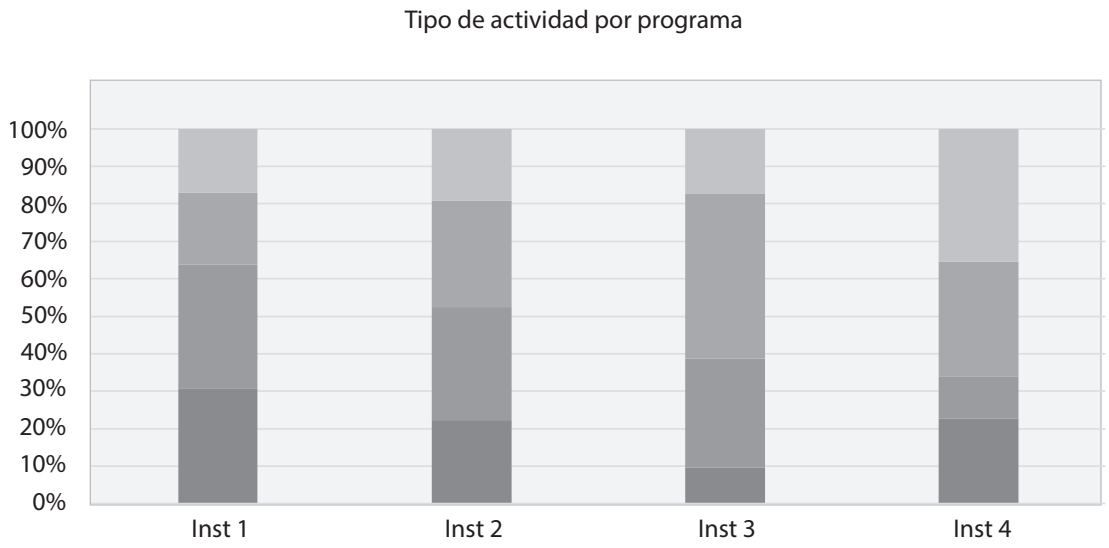


Figura 63. Barras apiladas al 100% del cruce entre actividad económica y programa seguido (institución)

La prueba y el análisis de los residuos estandarizados nos indicarán la medida de la significación de estas diferencias.

Se examinan los resultados de la prueba

En la tabla 92 se presentan la salida del IBM-SPSS para la prueba χ^2 de Pearson y para las medidas de asociación (tamaño del efecto) en esta prueba. Como se observa, la prueba muestra que la diferencia entre los grupos es ampliamente significativa (p<,001).

En cuanto a las medidas de tamaño del efecto, debe ignorarse la medida de Phi, ya que esta solo es apropiada para tablas 2x2. La *V* de Cramer, por su parte, indica un tamaño del efecto entre mediano y grande (0,13< V de Cramer=0,18 < 0,22).

Tabla 92. Salida del IBM-SPSS para la prueba Chi cuadrado de Pearson

Pruebas de Chi-cuadrado			
	Valor	**gl**	**Significación asintótica (bilateral)**
Chi-cuadrado de Pearson	34,863[a]	9	,000
Razón de verosimilitud	36,386	9	,000
Asociación lineal por lineal	14,186	1	,000
N de casos válidos	348		

a. 0 casillas (0,0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 13,01.

Medidas simétricas			
		Valor	**Significación aproximada**
Nominal por Nominal	Phi	,317	,000
	V de Cramer	,183	,000
N de casos válidos		348	

La tabla 93 presenta, de nuevo, el cruce entre las dos variables, con una fila más para cada institución: la correspondiente a los residuos estandarizados corregidos. Deben examinarse estos residuos e identificarse aquellos mayores a 1,96. Estas celdas son las que presentan diferencias a favor o en contra. Los valores absolutos mayores a 1,96 (≈2,00) se señalan en la tabla.

Tabla 93. Cruce entre actividad e institución con residuos estandarizados corregidos

			Actividad a un año				**Total**
			Estudia	**Trabaja**	**Estudia y trabaja**	**Ni est-Ni tra**	
Institución	**Inst 1**	**Recuento**	29	31	18	16	94
		Residuo corregido	2,7	1,4	-2,8	-1,1	
	Inst 2	**Recuento**	22	30	28	19	99
		Residuo corregido	,3	,8	-,6	-,5	
	Inst 3	**Recuento**	9	27	41	16	93
		Residuo corregido	-3,2	,4	3,3	-1,0	
	Inst 4	**Recuento**	14	7	19	22	62
		Residuo corregido	,3	-3,1	,0	3,1	
Total		**Recuento**	74	95	106	73	348

Los resultados indican que los egresados del programa cursado en la institución 1 muestran una mayor tendencia a continuar estudiando, pero una menor a estudiar y trabajar de forma simultánea. Los estudiantes de la institución 3, por el contrario, revelan una menor predisposición a continuar estudiando como actividad única, y una mayor tendencia a estudiar y trabajar simultáneamente. Los estudiantes egresados de la institución 4, por su parte, se distinguen por una menor actividad económica: trabajan en menor proporción y, en mayor proporción, no estudian ni trabajan.

Se expresan los resultados

Repetimos en este punto lo dicho sobre la misma prueba en el capítulo anterior. Para la expresión escrita de los resultados de la prueba χ^2 de Pearson, o sus similares, puede seguirse, en texto, el siguiente formato:

χ^2(<gl>)=<valor del Chi cuadrado> *p* =/< <valor de *p*> *V* = <valor V de Cramer>

O bien,

χ^2(<gl>, *N*=<tamaño de muestra>)=<valor del Chi cuadrado> *p* =/< <valor de *p*> *V* = <valor V de Cramer>

Siguiendo esta convención, los resultados pueden quedar expresados de la siguiente forma:

> *Se examinaron las diferencias en los niveles de actividad económica de los egresados de los programas educativos, a un año de concluido el programa, de las cuatro instituciones a través de una prueba de independencia χ^2 de Pearson. Los resultados indicaron diferencias significativas en la actividad económica entre los cuatro grupos de egresados con niveles de tamaño del efecto entre medianos y grandes $\chi^2(9,N=348)=34,86$ p < ,001 V = ,183.*
>
> *El análisis de los residuos estandarizados mostró que los egresados del programa en la institución 1 se distinguen por una mayor tendencia a estudiar y menor a estudiar y trabajar simultáneamente. Lo contrario ocurre con los egresados de la institución 3, que presentan menor propensión a estudiar y mayor a estudiar y trabajar de forma simultánea. Los egresados de la institución 4 revelan ser los menos activos, con menores intenciones de trabajar y mayores tendencias a no hacer lo uno ni lo otro (ni-ni).*

Pruebas para *k* medidas apareadas

En este punto se comprenden bien las diferencias entre las pruebas para comparar grupos independientes y la pruebas para comparar medidas apareadas o repetidas. En esta sección presentaremos y especificaremos las pruebas para examinar las diferencias entre *k* medidas repetidas.

Como sabemos, la decisión básica para la selección de la prueba específica depende del nivel de medida de las variables. En el primer caso, cuando las variables son métricas e intervalares, tenemos como candidata ideal la prueba del análisis de varianza de medidas repetidas (Anova MR).

Esta es una prueba paramétrica que tiene, como supuestos, la normalidad de las variables que representan cada una de las medidas repetidas y la esfericidad.

Como sabemos, para los Anova, el supuesto de normalidad no resulta ser muy estricto, sino que se requiere más bien una aproximación a la normalidad en la que no debería haber valores atípicos muy significativos. Si esta condición se cumple, deberemos pasar a examinar el segundo supuesto: el de esfericidad, que se refiere a la relativa igualdad de varianzas de las diferencias entre los niveles del factor de medidas repetidas. Si esta segunda condición también se cumple, podemos pasar al Anova MR sin correcciones. Si no se cumple, deberemos utilizar alguna de las correcciones disponibles para el procedimiento. Están disponibles dos correcciones: la corrección de Greenhouse-Geyser y la de Huynh-Feldt. Entre estas dos, se recomienda la última por cuanto ha mostrado ser más robusta.

Con en el caso de las pruebas para grupos independientes, en Anova MR es una prueba global que detecta diferencias entre las medidas repetidas pero no especifica entre cuáles. Para saberlo, necesitamos, de nuevo, del uso de pruebas *post hoc*. Usualmente se admiten las pruebas de Bonferroni y las de Holm, siendo la de Bonferroni la más conservadora.

Cuando el supuesto de normalidad es severamente violado o las medidas repetidas tienen un nivel ordinal, debemos recurrir a la prueba no paramétrica correspondiente al Anova MR: la prueba de medidas repetidas de Friedman, que no requiere de los supuestos de las paramétricas. Para el caso de que esta prueba muestre diferencias globales significativas, las alternativas de pruebas *post hoc* son más limitadas y se restringen al uso de la prueba *post hoc* de Connover.

Por último, debemos referirnos a la posibilidad de que las variables que representen las k medidas repetidas tengan un nivel estrictamente nominal. Para este caso presentaremos la prueba Q de Cochran. Esta prueba solo puede ser utilizada en el caso de que las medidas a examinar sean dicotómicas. Como pruebas *post hoc*, se recomendaría el uso de pruebas de McNemar sobre cada una de las medidas. Para el caso de medidas politómicas, se recomendaría el uso de regresiones logísticas de medidas repetidas, que excede ya el alcance de la presente obra.

El diagrama de la figura 64 presenta este árbol de decisiones. Iniciaremos con la presentación del caso de las variables métricas.

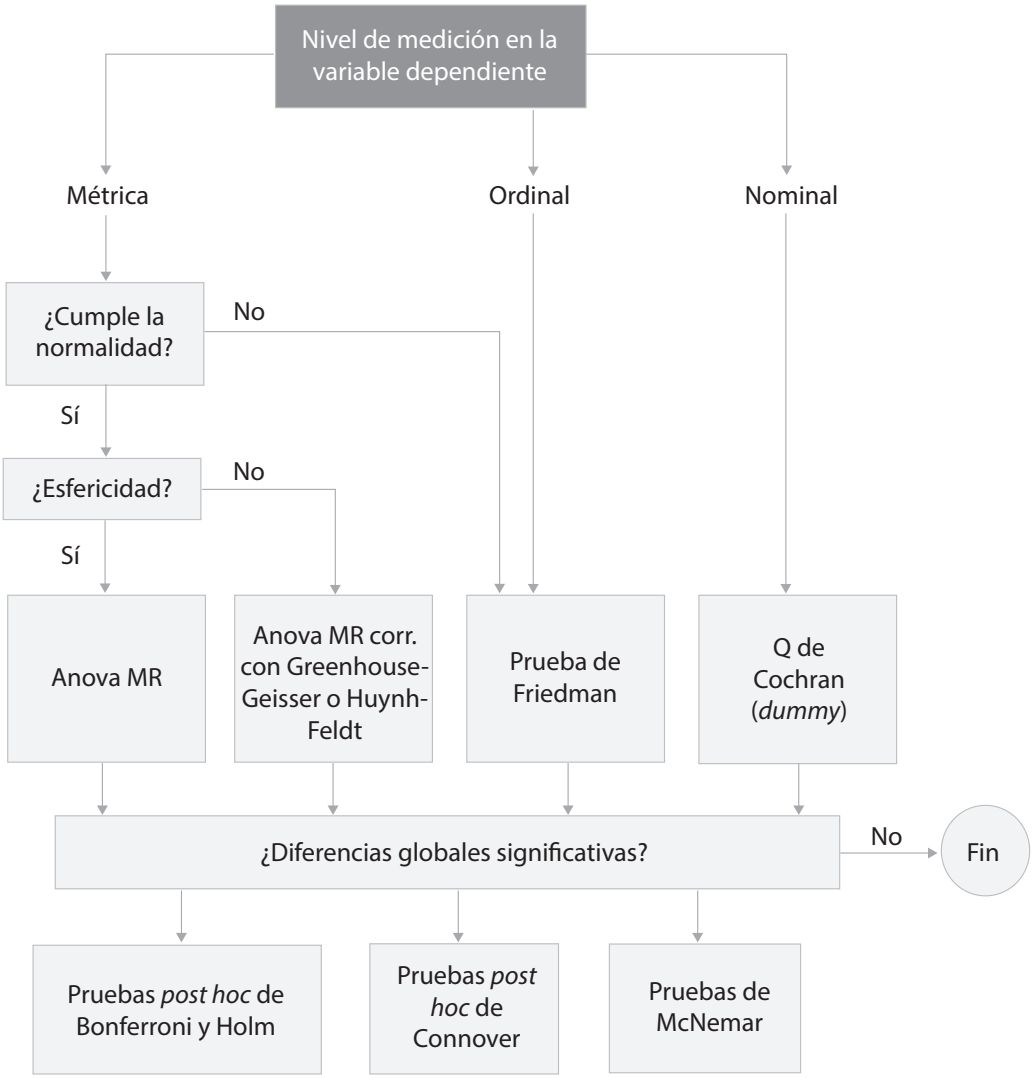


Figura 64. Diagrama de flujo para la decisión de pruebas para *k* medidas apareadas

Variable métrica: el Anova de medidas repetidas

Presentación

El *análisis de varianza de medidas repetidas (Anova MR)* es la prueba estadística que se usa para examinar si hay diferencias entre dos o más medidas tomadas para los mismos sujetos a lo largo del tiempo. Se utiliza para diseños de investigación intrasujeto, en los que queremos saber si hay cambios en una medida relacionados con el momento en que esta sea tomada.

La hipótesis nula de esta prueba es que no hay diferencias globales entre las *k* medidas. Por el contrario, la hipótesis alternativa indica que al menos una de las medidas presenta una media significativamente diferente de las otras.

El Anova MR requiere de varias variables dependientes métricas y continuas, que representan diferentes medidas del mismo constructo, hechas en diferentes momentos del tiempo sobre los mismos sujetos. Produce un estadístico (*F*), que representa la relación entre la varianza explicada y la varianza no explicada, una medida de los grados de libertad (gl) y un valor del nivel de significación (*p*) de la diferencia entre las medidas.

El Anova MR es una prueba paramétrica y, como tal, comparte los supuestos de la mayoría de las pruebas paramétricas; es decir, la normalidad de las variables dependientes y la igualdad de las varianzas, si bien ahora presenta algunas complejidades adicionales.

En cuanto al supuesto de normalidad, como en el caso del Anova de una vía, este supuesto no resulta ser demasiado estricto; básicamente se requiere que las diferentes medidas no presenten valores atípicos muy significativos.

En relación con el supuesto de igualdad de varianzas, para este caso, se conoce como *supuesto de esfericidad*, y consiste en la suposición de que debe haber igualdad en las varianzas de las diferencias entre todas las medidas repetidas. Si hay tres medidas repetidas, la esfericidad se da si hay igualdad entre las varianzas de las tres diferencias posibles (M1-M2, M1-M3 y M2-M3). Si fueran cuatro las medidas, entonces, la esfericidad supondría la igualdad entre las varianzas de las seis diferencias posibles (M1-M2, M1-M3, M1-M4, M2-M3, M2-M4 y M3-M4).

La esfericidad se evalúa mediante la aplicación del *test de esfericidad de Mauchly* (pronunciado "mockley"). Esta prueba examina la hipótesis nula de que las varianzas de todas las posibles diferencias entre las medidas son iguales. Si el valor del *p* obtenido por la prueba es mayor que 0,05, se asume esta condición. En el caso en que esto no pueda ser asumido, es necesario utilizar correcciones al estadístico *F*, a fin de no aumentar demasiado el error de tipo I.

Las correcciones más utilizadas cuando se viola el supuesto de esfericidad son las correcciones épsilon (ε) de Greenhouse-Geisser y de Huynh-Feldt. Se recomienda en este caso el uso de la corrección de Huynh-Feldt, en tanto ha mostrado ser más eficiente y robusta (Abdi, 2010).

En relación con las medidas de tamaño del efecto, el programa JASP ofrece cuatro posibilidades: el *eta cuadrado* (η^2), el *eta parcial al cuadrado* (η_p^2), el *eta al cuadrado general* (η_G^2) y el *omega cuadrado* (ω^2). El η^2 es una medida bastante popular, de fácil interpretación pero, en algunas ocasiones dificulta la comparación del efecto de la misma variable en distintos estudios, especialmente cuando hay varios factores independientes (que no es el caso en este momento), por lo que es preferible el uso del eta al cuadrado parcial (η_p^2), que resuelve este problema. Sin embargo, cuando las muestras son de tamaño pequeño ($n<30$), η_p^2 tiende a sesgarse, por lo que, en estos casos, se prefiere el uso del ω^2. El eta al cuadrado general (η_G^2) es particularmente útil cuando existen algunos factores controlados activamente por el investigador y otros individuales no controlados; no es este el caso.

La tabla 94 permite interpretar estas medidas de tamaño del efecto en el Anova MR.

Tabla 94. Límites para la interpretación de las medidas de tamaño del efecto en el Anova MR

Medida de tamaño del efecto	Nulo	Pequeño	Mediano	Grande
η^2	<0,1	0,1	0,25	0,37
η_p^2 (n>30)	<0,01	0,01	0,06	0,14
ω^2 (n<30)	<0,01	0,01	0,06	0,14

Como en el caso del Anova en una vía, el Anova MR es una prueba de carácter general que nos indica la presencia global de diferencias entre las medidas, pero no nos indica las medidas específicas que presentan diferencias entre sí. Para saberlo, es necesario examinar pruebas *post hoc* pero, como ya se ha dicho, estas pruebas únicamente pueden ser examinadas en el caso en el que la prueba global (el Anova MR) haya mostrado diferencias significativas.

La disponibilidad de pruebas *post hoc* en el Anova MR no es muy amplia. En el programa JASP están disponibles las pruebas *post hoc* de Holm y de Bonferroni. Esta última es un poco más conservadora que la primera. En el IBM-SPSS es posible incluir una de tres: DMS (que no representa ningún ajuste), Bonferroni y Sidak.

Cómo ejecutar un Anova de medidas repetidas

Para correr la prueba en el programa JASP, puede procederse a través del menú Anova (recuadro 42). En el programa IBM-SPSS debe buscarse el procedimiento a través del "modelo lineal general" (recuadro 43).

Recuadro 42. Cómo ejecutar un Anova MR en JASP

/ANOVA/Repeated Measures ANOVA… En este punto debe digitarse el nombre global de factor y los nombres de cada una de las mediciones en el cuadro "Repeated Measures Factor". Esto permitirá trasladar las variables de las mediciones al cuadro "Repeated Measures Cells"

Es recomendable seleccionar las siguientes opciones:

Display

√ Descriptive statistics

√ Estimates effect size

√ Partial η^2 √ ω^2 (dependiendo del tamaño de la muestra)

Assumption Checks

√ Sphericity test

(es posible aquí seleccionar correcciones, dependiendo del test anterior)

Sphericity corrections

√ None o √ Huynh-Feldt

Post Hoc test

En este punto se pasa el factor a la lista

Ö Bonferroni

Recuadro 43. Cómo ejecutar un Anova MR en IBM-SPSS

/Analizar/Modelo lineal general/Medidas repetidas…
En este punto debe digitarse el nombre global de factor (por defecto será "factor 1") y se debe definir el número de niveles (n) y pulsar el botón "Añadir" y el botón "Definir". Esto dará paso a otro menú "Medidas repetidas". Allí deben ser tomadas las variables de los n niveles y pasadas a "Variables intrasujetos".
• En el botón "Gráficos" puede pasar de la lista "Factores" a "Eje horizontal" y "Añadir"
Pulsar "Continuar"
• En el botón "Post hoc"… se selecciona la prueba adecuada
√ Tukey o √ Games-Howell
Pulsar "Continuar"
• En el botón "Opciones" es recomendable seleccionar:
√ Comparar los efectos principales. Allí arrastrar el nombre del factor a "Mostrar media para:" y seleccionar en la lista desplegable "Bonferroni".
√ Estadísticos descriptivos
√ Estimaciones del tamaño del efecto
Pulsar "Continuar"
Pulsar "Aceptar"

Es importante anotar que para el IBM-SPSS el Anova de medidas repetidas se establece como parte de un modelo mucho más general, conocido como el "modelo lineal general", que hace parte de la estadística multivariante. Por esta razón, el procedimiento arroja una serie de estadísticos multivariantes: la traza de Pillai, el lambda de Wilks, la traza de Hotelling y la raíz mayor de Roy. Estos estadísticos pueden ser ignorados en el caso de pruebas univariantes como las tratadas en el presente capítulo.

El ejemplo: variaciones en la vigilancia a lo largo del día

Desde el ámbito de la cronopsicología, se ha postulado la presencia de ciertos ritmos de funcionamiento cognitivo, de naturaleza circadiana (a lo largo del día) que podrían contribuir a un diseño más eficiente de la jornada escolar. Para explorar estos ritmos en población colombiana se planteó un proyecto de investigación que siguió un diseño de medidas repetidas a lo largo de la jornada.

El instrumento planteaba una tarea cognitiva sencilla, que requería de altas dosis de atención con un componente motor elevado. Se trataba de ordenar secuencias de siete dígitos durante un tiempo de cinco minutos. Esa tarea se planteaba a varios grupos de estudiantes al inicio, a la mitad y al finalizar la jornada escolar.

De esta tarea se obtuvieron dos indicadores: 1) uno de velocidad, que consiste en el número de secuencias abordadas para su ordenamiento, y 2) uno de eficiencia, que consiste en el porcentaje de secuencias correctamente ordenadas. Ambos indicadores son métricos y continuos.

Se pretende examinar si existen diferencias entre estos indicadores a medida que transcurre la jornada. La tarea fue resuelta por 149 estudiantes de grado noveno en dos instituciones educativas. Los datos provienen de un estudio efectivamente realizado y publicado (ver Hederich-Martínez *et al.*, 2005).

Planteamiento de las hipótesis

En esta ocasión, examinaremos la evolución de dos indicadores por medio del Anova MR. En el primero veremos si existen diferencias significativas en la *velocidad en la tarea*, indicada a través del número de ejercicios resueltos. Para este caso las hipótesis podrían quedar planteadas de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en las medias del número de ejercicios resueltos en los tres momentos de la jornada.*
>
> *Hipótesis alternativa (H_a). Existen diferencias significativas en cuanto a la media del número de ejercicios resueltos en, al menos uno, de los momentos de la jornada.*

En el segundo análisis examinaremos si hay diferencias en la *precisión* alcanzada, indicada por el porcentaje de ejercicios correctamente resueltos dentro del total de ejercicios. Al respecto, las hipótesis podrían ser formuladas de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en las medias del porcentaje de ejercicios correctamente resueltos en los tres momentos de la jornada.*
>
> *Hipótesis alternativa (H_a). Existen diferencias significativas en cuanto a las medias del porcentaje de ejercicios correctamente resueltos en, al menos, uno de los momentos de la jornada.*

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba*. El Anova MR tiene dos supuestos básicos: una relativa normalidad en las medidas y la esfericidad. El supuesto de esfericidad se examina a través del test de Mauchly. Los resultados de este test para las medidas de velocidad y las de precisión aparecen en la tabla 95.

Tabla 95. Test de esfericidad de Mauchly para las medidas de velocidad y precisión

	ω de Mauchly	**Approx. X^2**	**gl**	**p**
Velocidad	0,952	5,798	2	,055
Precisión	0,777	29,581	2	< .001

De acuerdo con la tabla, el test indica que es posible asumir la esfericidad para el caso de las medidas de velocidad $W(2)=0{,}952$ $p=,055$, pero no se puede hacer esa suposición para el caso de las tres medidas de precisión $W(2)=29{,}58$ $p<,001$. Esto nos indica que podemos correr el Anova de medidas repetidas convencional para las medidas de velocidad, pero no podemos hacerlo para las medidas de precisión, en donde necesitaremos utilizar la corrección de Huynh-Feldt.

- *Se examinan resultados descriptivos.* Las gráficas y la tabla muestran los estadísticos descriptivos de cada una de las mediciones. Como se observa, las medidas de velocidad inician la jornada en su punto más bajo; para la segunda toma se advierte un incremento bastante pronunciado en el número de ejercicios resueltos; y para la tercera, esto se incrementa aún más, aunque ahora el incremento es más bien leve.

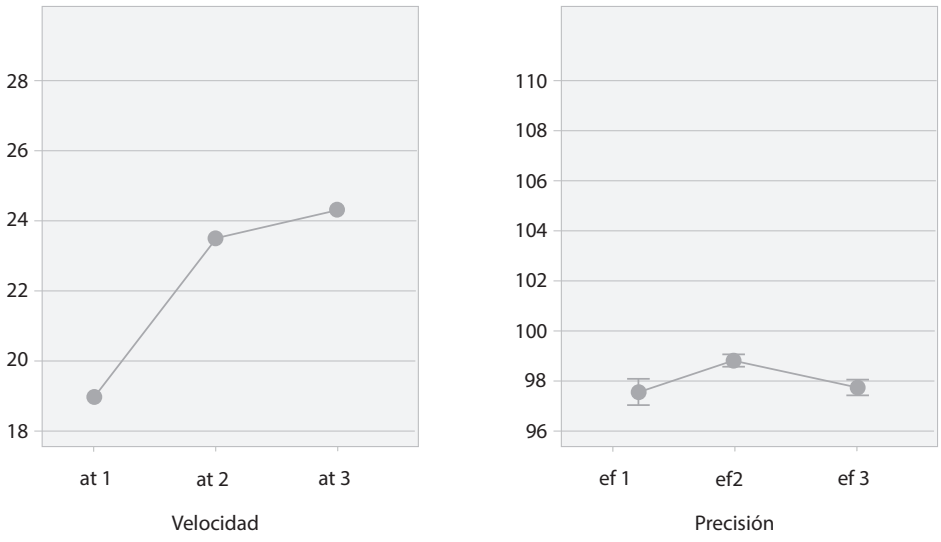


Figura 65. Medias de velocidad y precisión a lo largo de la jornada

Tabla 96. Descriptivos de velocidad y precisión a lo largo de la jornada

	Velocidad		Precisión	
	M	DE	M	DE
Primera toma	18,97	6,03	97,53	5,78
Segunda toma	23,49	6,08	98,78	2,63
Tercera toma	24,32	6,31	97,72	3,23

Para el caso de las medidas de precisión, indicadas por el porcentaje de ejercicios correctamente resueltos dentro del total de ejercicios abordados, se observa que este indicador no se distancia demasiado de una precisión del 100 %. La primera toma muestra los menores niveles de precisión de las tres; estos niveles ascienden para la segunda y vuelven a descender para la tercera. La prueba global y, si procede, las pruebas *post hoc* nos indicarán el nivel de significación de estas diferencias.

Se examinan los resultados de la prueba

Las tablas presentan el resultado de la prueba Anova de medidas repetidas sobre los tres indicadores de velocidad. Como se observa, la prueba muestra diferencias muy significativas entre las tres medidas de velocidad (p<,001), con un tamaño del efecto que podría ser considerado grande ($\eta^2_p = 0{,}470 > 0{,}14$).

El resultado significativo del Anova MR nos autoriza a correr las pruebas *post hoc* para el examen de las diferencias entre las aplicaciones. Los resultados de los niveles de significación con las correcciones de Bonferroni y Holm se presentan en la tabla 97.

Tabla 97. Tabla de resultados del Anova MR para la velocidad por momento de la jornada

Casos	Suma de cuadrados	gl	Media cuadrática	F	p	η^2	η^2_p	η^2_G	ω^2
Velocidad	1974,829	2	987,415	104,850	< ,001	0,470	0,470	0,129	0,127
Residuals	2222,504	236	9,417						

Nota: suma de cuadrados tipo III.

Tabla 98. Pruebas post hoc de velocidad entre diferentes momentos de la jornada

Comparaciones *post hoc*–Velocidad							
		Diferencia media	SE	t	d de Cohen	p_{bonf}	p_{holm}
at1	at2	-4,521	0,398	-11,364	-1,042	< ,001 ***	< ,001 ***
	at3	-5,353	0,398	-13,455	-1,233	< ,001 ***	< ,001 ***
at2	at3	-0,832	0,398	-2,091	-0,192	,113	,038 *

* p < .05, *** p < .001

Notas: la *d* de Cohen no corrige las comparaciones múltiples; el valor de *p* fue ajustado para comparar una familia de 3.

De acuerdo con la tabla, las pruebas *post hoc* indican diferencias muy significativas en los niveles de velocidad entre todas las aplicaciones, que resultan ser muy significativos entre la primera y la segunda aplicación, así como entre la primera y la tercera (p<,001). Las diferencias entre la segunda y la tercera aplicación manifiestan también ser significativas en la prueba *post hoc* con corrección de Holm, aunque en niveles menores (*p*=,038), pero no alcanza los niveles de significación esperados en la prueba *post hoc* con la corrección de Bonferroni (p=,113), que resulta más conservadora. En ese sentido, los resultados indican que la velocidad en la ejecución de la tarea aumenta considerablemente a lo largo de la jornada, encontrando un máximo hacia el final de esta.

En relación con las diferencias en los niveles de precisión, la tabla muestra los resultados del análisis de varianza de MR, con las correcciones solicitadas por el incumplimiento del supuesto de esfericidad. Como se observa, el programa JASP nos recuerda, en la tabla 99, que se violó el supuesto de esfericidad. Para nuestro caso, y dados los niveles de épsilon encontrados, se recomienda el uso de la corrección de Huynh-Feldt.

Tabla 99. Tabla de resultados del Anova mr con, y sin, correcciones para la precisión por momento de la jornada

Dentro de los efectos de los sujetos								
Casos	**Corrección de esfericidad**	**Suma de cuadrados**	**gl**	**Media cuadrática**	**F**	**p**	**η^2**	**η^2_p**
Precisión	Ninguno	107.393 [a]	2.000 [a]	53.696 [a]	3.385 [a]	0.036 [a]	0.028	0.028
	Greenhouse-Geisser	107.393	1.635	65.692	3.385	0.045	0.028	0.028
	Huynh-Feldt	107.393	1.655	64.905	3.385	0.045	0.028	0.028
Residuos	Ninguno	3744.190	236.000	15.865				
	Greenhouse-Geisser	3744.190	192.905	19.410				
	Huynh-Feldt	3744.190	195.245	19.177				

Nota: suma de los cuadrados tipo III.

a El test de esfericidad de Mauchly indica que se viola la suposiciòn de esfericidad (p < .05).

Los resultados indican que el análisis de varianza con la corrección de Huynh-Feldt muestra diferencias globales significativas entre las tres medidas de precisión en la ejecución de la tarea (*p*=,045), y las medidas de tamaño del efecto muestran ser entre pequeñas y medianas ($0{,}01 < \eta^2_p = 0{,}028 < 0{,}06$).

Puede ser importante observar que aunque la prueba global mostró ser significativa, los niveles de significación se encuentran ya muy cerca del umbral de decisión, así como las estimaciones de tamaño del efecto. Por otra parte, vale la pena notar que la comparación entre el Anova MR sin y con corrección muestra únicamente diferencias en los grados de libertad y en los niveles de significación, que con las correcciones aumentan levemente. Las estimaciones de tamaño del efecto son iguales en los dos casos.

La presencia de diferencias significativas nos autoriza, de nuevo, al examen de las pruebas *post hoc* entre las aplicaciones. Los resultados se presentan en la tabla 100 e indican la presencia de diferencias significativas entre los niveles de precisión en la primera y la segunda toma, con las dos correcciones examinadas (*p*=,050). Las diferencias no alcanzan a ser significativas entre la primera y la segunda toma, ni entre la segunda y la tercera toma en ninguna de las pruebas examinadas. Aparentemente, los niveles de precisión alcanzados son significativamente mayores a la mitad de la jornada en comparación con el inicio.

Tabla 100. Pruebas **post hoc** *de Bonferroni y Holm para las diferencias en la precisión a lo largo de la jornada*

Comparaciones *post hoc*–Precisión							
		gl	**SE**	**t**	**d de Cohen**	**p_{bonf}**	**p_{holm}**
ef1	ef2	-1.245	0.516	-2.412	-0.221	0.050 *	0.050 *
	ef3	-0.186	0.516	-0.361	-0.033	1.000	0.719
ef2	ef3	1.059	0.516	2.051	0.188	0.124	0.083

* p < .05

Notas: la d de Cohen no corrige las comparaciones múltiples; el valor de p fue ajustado para comparar una familia de 3.

Se expresan los resultados

Para expresar en texto los resultados del Anova MR puede seguirse el siguiente formato:

$$F(<gl1>, <gl2>)= <valor\ F> \ p </= <valor\ p> \ \eta_p^2 = <valor\ \eta_p^2>$$

Utilizando este formato, los resultados podrían ser descritos como sigue:

> *Se utilizó un Anova de medidas repetidas para el examen de las diferencias en la velocidad de ejecución entre los tres momentos de la jornada: al inicio, en un momento intermedio y al final de la jornada. La prueba de Mauchly indicó que podía ser asumido el supuesto de esfericidad W(2)=0,952 p=,055 y el Anova MR mostró diferencias globales significativas con un tamaño del efecto que podría ser considerado grande F(2,236)=104,85 p<,001* η^2_p *= 0,470. Los resultados de las pruebas post hoc de Holm indicaron que el número de ejercicios resueltos muestra un mínimo al inicio de la jornada (M=18,98 DE=6,03), mientras que en el segundo momento este número se incrementa de forma importante (M=23,50 DE=6,09) y muy significativa (p<,001). Para el tercer momento de la jornada, el número de ejercicios vuelve a mostrar un incremento importante en promedio (M=24,33 DE=6,31) que resulta ser significativo frente al primero y frente al segundo momento (p<,001 y p=,038, respectivamente).*
>
> *Por otra parte, se utilizó un Anova MR para el examen de las diferencias en los niveles de precisión en la respuesta en los tres momentos de la jornada. Para este caso, la prueba de Mauchly indicó que no podía asumirse el supuesto de esfericidad W(2)=29,58 p<,001, por lo que debió correrse el Anova de medidas repetidas utilizando la corrección de Huynh-Feldt por la violación del supuesto. Los resultados de esta prueba revelaron diferencias significativas en la precisión mostrada por los estudiantes entre las aplicaciones con un tamaño del efecto entre pequeño y mediano F(1,66, 195,23)=3,39 p=,045* η^2_p *=0,028. Las pruebas post hoc de Holm exponen un valor del indicador de precisión a la mitad de la jornada (M=98,78 DE=2,64) que resulta ser significativamente mayor (p=0,05) al mostrado al inicio de la jornada (M=97,54 DE=5,79). Al final de la jornada, el indicador de precisión vuelve a bajar a niveles similares de los del inicio (M=97,72 DE=3,24), sin que se verifiquen diferencias significativas entre este momento y el primero o el segundo (p=,719 y p=,083, respectivamente).*

Variable ordinal: la prueba de Friedman

Presentación

En las situaciones en las que tenemos una serie de dos o más medidas repetidas pero no podemos utilizar un Anova de medidas repetidas, ya sea porque se incumplió seriamente el supuesto de normalidad o porque nuestras variables dependientes tienen un nivel de medida ordinal, debemos recurrir al uso del equivalente no paramétrico del Anova MR: la prueba de Friedman, también conocida como Anova de medidas repetidas de Friedman.

La *prueba de Friedman* es, como todos los equivalentes no paramétricos que hemos examinado hasta ahora, una prueba basada en rangos que resulta ser libre de distribución. Por esta razón no requiere del cumplimiento de los supuestos de las pruebas paramétricas. Solo necesita de los aspectos básicos del diseño de investigación *intrasujetos*: una serie de varias mediciones con un nivel de medida, al menos, ordinal, efectuadas sobre la misma muestra de sujetos.

Cuando solo se dispone de dos mediciones, la prueba de Friedman es equivalente a la prueba de los rangos con signo de Wilcoxon para el examen de dos medidas apareadas que presentamos en el capítulo anterior; de hecho, la prueba de Friedman puede ser considerada como una extensión de esta última.

La prueba de Friedman presenta un valor de Chi-cuadrado, una medida de los grados de libertad (gl) y un valor de probabilidad asociado con el Chi-cuadrado (p).

Como en el caso de todas las pruebas que hemos examinado en el presente capítulo, la prueba de Friedman da cuenta de diferencias globales entre las mediciones, sin indicar las mediciones específicas que presentan diferencias significativas entre sí. En el caso de que la prueba nos muestre diferencias globales significativas deberemos correr pruebas *post hoc* para examinar las mediciones específicas que presentan diferencias.

No hay muchas pruebas *post hoc* apropiadas para esta situación. En el programa JASP se aportan la *prueba post hoc de Connover*, que representan pruebas no paramétricas apropiadas a la situación y se aportan también las correcciones del nivel de significación de Bonferroni y Holm. En el IBM-SPSS no se ofrece ninguna posibilidad de pruebas *post hoc* apropiadas para la prueba de Friedman.

Al respecto de la medición del tamaño del efecto observado en la prueba de Friedman, debe anotarse que, tal y como ocurre con la prueba de Kruskal-Wallis, no hay una manera general y consensualmente aceptada de calcular el tamaño del efecto en esta prueba. Tomczak y Tomczak (2014) sugieren el uso de un estadístico conocido como la *W de Kendall*, también conocido como el *coeficiente de concordancia de Kendall*. Este es una normalización de la estadística de la prueba de Friedman que varía entre 0 (no hay acuerdo/efecto) y 1 (acuerdo/efecto total). El coeficiente de concordancia de Kendall es entregado, por defecto, tanto en el JASP como en el IBM-SPSS cuando se corre la prueba de Friedman.

Ejecutar la prueba de Friedman

Para correr la prueba de Friedman en el *software* JASP debe procederse a través del menú de Anova (recuadro 44). En el programa IBM-SPSS debe buscarse el procedimiento a través de "Pruebas no paramétricas" (recuadro 45).

Recuadro 44. Cómo ejecutar la prueba de Friedman en JASP

/ANOVA/Repeated measures ANOVA… En este punto debe digitarse el nombre global de factor y los nombres de cada una de las mediciones en el cuadro "Repeated Measures Factor". Esto permitirá trasladar las variables de las mediciones al cuadro "Repeated Measures Cells"

Es recomendable seleccionar las siguientes opciones:

Display

√ Descriptive statistics

Descriptive Plots (pasar la variable de factor a "Horizontal Axis")

√ Display error bars

Nonparametrics (pasar la variable de Factor a "RM Factor")

√ Connover post hoc test

Recuadro 45. Cómo ejecutar la prueba de Friedman en IBM-SPSS

/Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/K muestras relacionadas…
En este punto se deben arrastrar las K variables a la lista "Variables de prueba".
√ Friedman
Pulsar el botón "Aceptar"

El ejemplo: variaciones en la memoria a corto plazo a lo largo del día

En el mismo estudio cronopsicológico, una de cuyas partes fue expuesta para ejemplificar el uso del Anova MR, fue examinado el funcionamiento de la memoria de corto plazo en los estudiantes colombianos. Específicamente, fueron examinados los procesos de reconocimiento inmediato de proposiciones en tres momentos de la jornada.

Para hacerlo, fue leído un breve texto, cuya lectura tomó cinco minutos, e inmediatamente después, fue administrado un cuestionario a los estudiantes con treinta preguntas, para ser respondidas "sí" o "no" con proposiciones que podían o no estar presentes en el texto leído.

Los modelos teóricos del funcionamiento de la memoria humana han constatado la presencia de dos procesos funcionalmente diferentes: el reconocimiento de proposiciones efectivamente presentes en el texto y la identificación de proposiciones ausentes en este. Examinaremos la evolución a lo largo de la jornada de estos dos procesos.

En total contamos con tres mediciones: al inicio, en un momento intermedio y al final de la jornada. En cada una de ellas se leyó un texto y se identificó, mediante la aplicación de un cuestionario, la eficiencia en: 1) el reconocimiento de proposiciones presentes y 2) la identificación proposiciones ausentes. Cada una de las medidas de eficiencia en el reconocimiento se codificó en una escala ordinal de tres puntos, a saber: 1) baja eficiencia, 2) eficiencia media y 3) alta eficiencia.

Contamos con información de 96 estudiantes colombianos en dos instituciones educativas y dos diferentes jornadas escolares. Los datos presentados provienen de una observación real previamente publicada, si bien los datos mismos fueron recodificados con propósitos pedagógicos (véase Hederich-Martínez *et al.*, 2005).

Planteamiento de las hipótesis

La prueba de Friedman examina las diferencias en las distribuciones y las medianas de las diferentes mediciones. En este caso examinaremos dos pruebas diferentes: una para el reconocimiento de proposiciones presentes y otra para la identificación de proposiciones ausentes. Iniciando con la primera, las hipótesis pueden ser:

> *Hipótesis nula (H_0). No hay diferencias en las distribuciones de la eficiencia en el reconocimiento de proposiciones presentes a lo largo de los tres momentos de la jornada.*
>
> *Hipótesis alternativa (H_1). Al menos una de las mediciones de eficiencia en el reconocimiento de proposiciones presentes es diferente de las otras mediciones hechas a lo largo de la jornada.*

Formulando las mismas hipótesis en términos de la eficiencia en la identificación de proposiciones ausentes, quedarían de la siguiente forma:

> *Hipótesis nula (H_0). No hay diferencias en las distribuciones de la eficiencia en la identificación de proposiciones ausentes a lo largo de los tres momentos de la jornada.*
>
> *Hipótesis alternativa (H_1). Al menos una de las mediciones de la eficiencia en la identificación de proposiciones ausentes es diferente de las otras mediciones hechas a lo largo de la jornada.*

Se corre la prueba

- *Se examinan resultados descriptivos*. En nuestro caso tenemos dos grupos de mediciones ordinales de tres puntos. Formalmente, para examinar diferencias entre estas mediciones, debemos desplegar las frecuencias de cada una de ellas, tal y como aparecen en la tabla 101, y luego graficarlas como barras apiladas al 100 %, como aparece en la figura 66.

Tabla 101. Cruce entre la eficiencia en el reconocimiento de proposiciones presentes e identificación de proposiciones. Ausentes por momento de la jornada

	Reconocimiento de prop. presentes			Identificación de prop. ausentes		
	Inicio	**Intermedio**	**Fin**	**Inicio**	**Intermedio**	**Fin**
Bajo	2	1	3	27	12	43
Medio	42	56	57	56	59	46
Alto	52	39	36	13	25	7
Total	96	96	96	96	96	96

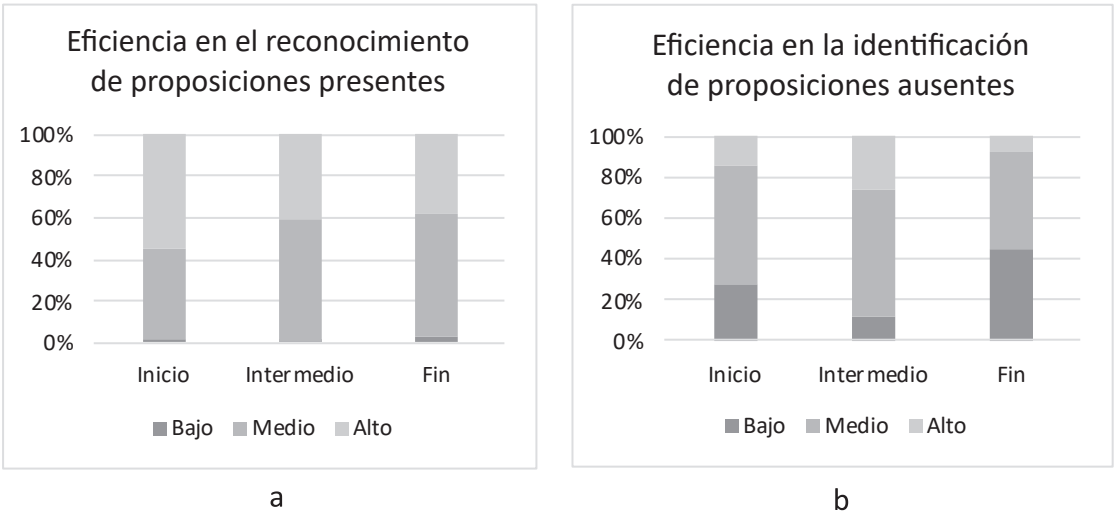


Figura 66. Gráficas de barras apiladas al 100 % de eficiencia en el reconocimiento e identificación de proposiciones por momento de la jornada

El examen de las gráficas indica que, frente al reconocimiento de proposiciones presentes, el área de las barras que representa una alta eficiencia inicia con casi el 60 % del total y va disminuyendo

progresivamente con cada medición; se observa un salto especialmente grande entre la primera medición (inicio) y la segunda (intermedio).

En relación con la eficiencia en la identificación de proposiciones ausentes, la inspección de la gráfica muestra, en un contexto general con menores niveles de eficiencia, que la primera medición inicia con niveles bajos de eficiencia, que se incrementan bastante en la segunda y vuelven a bajar para la tercera.

Las gráficas muestran las tendencias que hemos descrito, hechas a partir de las medias entre las diferentes mediciones. Para este caso, el reconocimiento de proposiciones presentes aparece como "MCP verdaderas" y la identificación de proposiciones ausentes como "MCP falsas". Sobra decir que, formalmente, no deben ser presentadas estas gráficas, debido a que, al trabajar con variables ordinales, no estamos autorizados a calcular medias de estas variables.

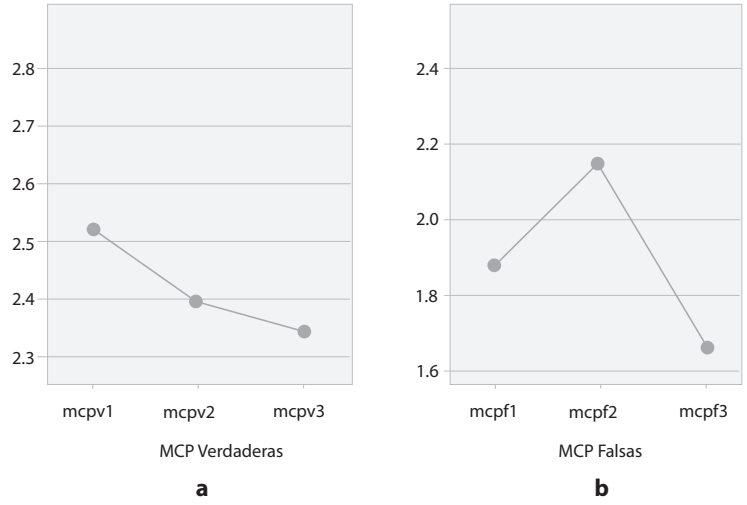


Figura 67. Medias de la eficiencia en el reconocimiento e identificación de proposiciones por momento de la jornada

En general, el comportamiento de los dos grupos de mediciones muestra ser bastante diferente. El reconocimiento de proposiciones efectivamente presentes inicia con máxima eficiencia al principio de la jornada, manifiesta un brusco descenso en el momento intermedio y llega a su mínimo al final de la jornada. Por su parte, la eficiencia identificación de proposiciones ausentes inicia con niveles de eficiencia medio-bajos, llega a su máximo en el segundo momento de la jornada y vuelve a descender bruscamente al final de esta. Las pruebas nos indicarán la magnitud de diferencias específicas.

- *Se examinan los supuestos y se selecciona la prueba.* Los datos cumplen con todos los supuestos para el examen de las pruebas de Friedman. En los dos casos tenemos tres mediciones ordinales efectuadas sobre una muestra de 96 sujetos. La prueba no requiere más supuestos que estos.

Se examinan los resultados de la prueba

El resultado de la prueba de Friedman para las tres mediciones de la eficiencia en el reconocimiento de proposiciones presentes se presenta en la tabla 102. Se evidencia que la prueba muestra diferencias globales significativas entre las tres mediciones (p=,035). La W de Kendall encontrada

(,495) indicaría una relación media entre las mediciones. Interpretada como medida de tamaño del efecto, asumimos que representa un tamaño medio del efecto.

Tabla 102. Resultados de la prueba de Friedman de reconocimiento de proposiciones presentes a lo largo de la jornada

Prueba de Friedman				
Factor	**Chi cuadrado**	**gl**	**p**	**ω de Kendall**
MCP verdaderas	6.700	2	0.035	0.495

En la medida en que la prueba de Friedman indica diferencias globales significativas, podemos proceder al examen de las diferencias, a través de las pruebas *post hoc* de Connover. Los resultados se presentan en la tabla 102.

Tabla 103. Pruebas post hoc de reconocimiento de proposiciones presentes a lo largo de la jornada

Comparaciones *post hoc* de Conover–MCP verdaderas								
		T-Stat	**gl**	**W_i**	**W_j**	**p**	**p_{bonf}**	**p_{holm}**
mcpv1	mcpv2	1.750	190	205.500	189.000	0.082	0.245	0.164
	mcpv3	2.545	190	205.500	181.500	0.012	0.035	0.035
mcpv2	mcpv3	0.795	190	189.000	181.500	0.427	1.000	0.427

Como se observa, las pruebas *post hoc* indican que entre la primera medición (mcpv1) y la segunda (mcpv2) la diferencia no alcanza a ser significativa en los niveles convencionalmente aceptados, aunque no está lejos del límite (p=0,082). Ya para la tercera medición (mcpv3) los niveles de eficiencia bajan aún más y se alcanzan a encontrar diferencias significativas con la primera medición, ya sea que miremos la prueba sin correcciones (p=,012) o con las correcciones de Boferroni y Holm (p=,035 en los dos casos). No hay diferencias significativas entre la segunda y la tercera medición.

Ahora, por el lado del examen de las diferencias en cuanto a la identificación de proposiciones ausentes, la prueba de Friedman indica diferencias globales muy significativas (p<,001) con un grado de relación entre las mediciones medio (W=,540). Las pruebas p*ost hoc* de Connover, por su parte, exponen diferencias muy significativas entre todas las mediciones, pero la diferencia más significativa se presenta entre la segunda y la tercera $t(190)=6{,}25$ p<,001 (véase tablas 104 y 105).

Tabla 104. Resultados de la prueba de Friedman de identificación de proposiciones ausentes a lo largo de la jornada

Prueba de Friedman				
Factor	**Chi cuadrado**	**gl**	**p**	**ω de Kendall**
MCP Falsas	38.956	2	< .001	0.540

Tabla 105. Pruebas post hoc de identificación de proposiciones ausentes a lo largo de la jornada

Comparaciones *post hoc* de Conover–MCP falsas								
		T-Stat	gl	W_i	W_j	p	p_{bonf}	p_{holm}
mcpf1	mcpf2	3.478	190	189.500	226.500	< .001	0.002	0.001
	mcpf3	2.773	190	189.500	160.000	0.006	0.018	0.006
mcpf2	mcpf3	6.251	190	226.500	160.000	< .001	< .001	< .001

Se expresan los resultados

Los resultados de la prueba de Friedman pueden ser indicados en texto mediante el siguiente formato:

$$\chi^2 (<gl>)= <Valor > p</= <Valor p>$$

Utilizando este formato, los resultados podrían ser descritos como sigue:

> *Para examinar las diferencias en la eficiencia en el reconocimiento de proposiciones presentes y en la identificación de proposiciones ausentes entre los tres momentos de la jornada (al inicio, en un punto intermedio y al final) se utilizó la prueba no paramétrica de Friedman.*
>
> *Iniciando con el reconocimiento de proposiciones presentes, la prueba mostró diferencias globales significativas $\chi^2(2) = 6,70$ p = ,035. Los resultados de la prueba post hoc de Connover indican que los estudiantes inician la jornada con un máximo de eficiencia, que disminuye de forma apreciable para el momento intermedio, sin que se presenten diferencias significativas con el primero (p=,082); para el tercer momento el nivel de la eficiencia desciende al mínimo, y alcanza a registrar ahora diferencias significativas con el inicio (p=,012), pero no con el momento anterior (p=,427).*
>
> *En cuanto a la eficiencia en la identificación de proposiciones ausentes, la prueba de Friedman mostró diferencias globales significativas $\chi^2(2) = 38,96$ p < ,001. Las pruebas* post hoc *de Connover indicaron que el nivel de eficiencia inicia en un punto medio y se incrementa, de forma muy significativa, para el segundo momento de la jornada (p<,001); a partir de allí, el nivel de eficiencia desciende hasta el mínimo en el momento final de la jornada, con lo cual presenta diferencias significativas con el anterior y con el inicio (p<,001 y p=,006, respectivamente).*

Variable nominal: la Q de Cochran

Presentación

Finalmente, hemos llegado a la situación en la que tenemos un diseño intrasujeto, con más de dos medidas, estrictamente nominales y apareadas, y queremos conocer si existen cambios entre las diferentes mediciones.

Esta situación se presenta pocas veces en la investigación educativa. En primer lugar, porque la mayoría de los investigadores prefieren utilizar medidas numéricas para verificar los cambios en

los diseños intrasujeto y, cuando ello no es posible, optan por el uso de medidas ordinales. En segundo lugar, porque los diseños de medidas repetidas, y mucho más si incluyen tres o más medidas, tienden a ser largos, y por tanto muy costosos.

Las pruebas para el examen de esta situación dependen de la cantidad de valores que presentan las medidas repetidas. Cuando las medidas son nominales dicotómicas puede ser utilizada la prueba Q de Cochran. Si, por el contrario, las k medidas son nominales politómicas, es necesario utilizar un procedimiento por entero diferente: la regresión logística de medidas repetidas. Este procedimiento resulta extremadamente complejo para el nivel, por lo que tendrá que ser explicado mucho más adelante en una publicación especializada.

La *prueba Q de Cochran* (pronunciado "Cocran") es una prueba estadística no paramétrica que constituye la generalización de la prueba de McNemar. Como se explicó en el capítulo 10, utilizamos la prueba de McNemar para el examen de cambios intrasujeto en dos medidas nominales dicotómicas apareadas; ahora la prueba Q de Cochran considera la presencia de k (k>2) medidas dicotómicas apareadas.

La prueba Q de Cochran produce un estadístico (Q), una medida de los grados de libertad (gl) que depende de la cantidad de mediciones y un valor de probabilidad asociado con el estadístico (*p*).

Como prueba *post hoc* a la prueba Q de Cochran, sugerimos el uso de la prueba de McNemar.

Respecto de las mediciones de tamaño del efecto, no encontramos propuestas directas. Sugerimos, como en el caso de la prueba de Friedman, y en gracia a las similitudes entre estas dos pruebas, el uso del coeficiente de concordancia W de Kendall como medida de tamaño del efecto.

Ejecutar la prueba Q de Cochran

El cálculo de la prueba Q de Cochran no es ofrecido como una posibilidad en el programa JASP. Para examinar esta prueba en el IBM-SPSS, debe buscarse el procedimiento a través de "Pruebas no paramétricas":

Recuadro 46. Cómo ejecutar la prueba Q de Cochran en IBM-SPSS

/Analizar/Pruebas no paramétricas/Cuadros de diálogo antiguos/K muestras relacionadas…
En este punto se deben arrastrar las k variables a la lista "Variables de prueba:"
√ W de Kendall
√ Q de Cochran

El programa IBM-SPSS presentará muy poca información. Al respecto de la Q de Cochran estarán las frecuencias de las variables elegidas, los grados de libertad y los valores de probabilidad. Sobre la W de Kendall, ofrecerá el valor de W y un valor de Chi-cuadrado igual al de la Q ya expresado.

El ejemplo: variaciones en una opinión a lo largo de cuatro años

En un programa de seguimiento integral a las instituciones educativas públicas de nivel secundario de la ciudad de Cali, Colombia, llevado a cabo durante cuatro años consecutivos, se realizó

una inversión importante en infraestructura y apoyo pedagógico y administrativo en un grupo de instituciones educativas. Como parte del proceso de evaluación de impacto de la intervención, se hicieron diferentes mediciones sucesivas de clima escolar. Entre estas mediciones, se preguntó a los docentes de uno de los planteles por sus relaciones con sus directivos docentes. Concretamente, se pidió a los profesores que respondieran con "sí" o "no" a la pregunta de si en la institución existen buenas relaciones entre docentes y directivos docentes.

Contamos con las respuestas de cuarenta docentes durante cuatro años consecutivos. En cada caso se codificó como "0" una respuesta que indicaba que no existían buenas relaciones, y como "1" una respuesta que indicaba que sí existían buenas relaciones.

Planteamiento de las hipótesis

Las hipótesis se pueden formular como sigue:

> *Hipótesis nula (H_0). No hay diferencia entre las cuatro mediciones acerca de la opinión sobre las relaciones entre docentes y directivos docentes.*
>
> *Hipótesis alternativa (H_1). Al menos una de las mediciones de opinión sobre las relaciones entre docentes y directivos docentes es diferente de las otras.*

Se corre la prueba

- *Se examinan los supuestos y se selecciona la prueba.* La prueba Q de Cochran requiere un diseño de k medidas apareadas, cada una de las cuales debe tener un nivel de medida nominal dicotómica. En nuestro caso, tenemos cuatro medidas dicotómicas, cada una correspondiente a un año, del 2016 al 2019. Se cumplen los supuestos para la aplicación de la prueba Q de Cochran.
- *Se examinan resultados descriptivos.* La tabla 106 muestra los resultados de las frecuencias de cada una de las cuatro medidas repetidas. Estos se encuentran representados gráficamente en la figura 68, en términos de porcentajes. Como se observa, la curva muestra un comportamiento creciente entre los cuatro años que inicia en el 2016 con poco más del 50 % de los docentes que responden "sí" a la pregunta sobre relaciones con los directivos, y concluye en el 2019 con un valor cercano al 70 % de docentes que responden afirmativamente a la misma pregunta. La medida de la significación de esta diferencia tendrá que ser respondida por la prueba.

Tabla 106. Frecuencias de cada medición

Frecuencias		
	Valor	
	0	1
a2016	19	21
a2017	15	25
a2018	16	24
a2019	12	28

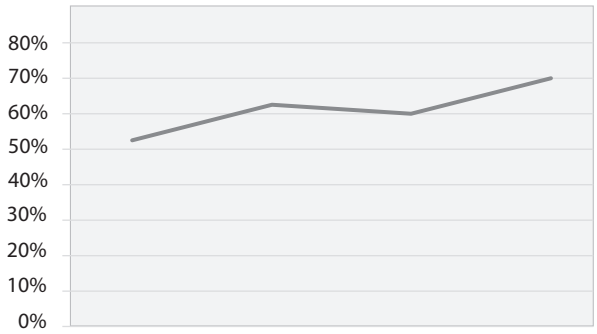


Figura 68. Porcentaje de docentes con respuesta "sí" por año

Se examinan los resultados de la prueba

La tabla 107 presenta los resultados de la prueba. Como se observa, la aplicación de la prueba muestra que hay diferencias globales significativas al respecto del porcentaje de docentes que, durante cuatro años consecutivos, responden "sí" a la pregunta de si tienen buenas relaciones con los directivos.

Tabla 107. Resultados de la prueba Q de Cochran

N	40
Q de Cochran	9,375[a]
gl	3
Sig. asintótica	,025

a. 0 se trata como un éxito

La tabla 108 presenta los resultados de la prueba W de Kendall para estas medidas repetidas. Si utilizamos los resultados de esta prueba como medida de tamaño de efecto, y lo asimilamos a la interpretación que hace Cohen al coeficiente *r* de Pearson, diríamos que un *W*=,078 = ,08 indicaría un tamaño del efecto pequeño.

Tabla 108. Resultados de la prueba W de Kendall

Estadísticos de prueba	
N	40
W de Kendall[a]	,078
Chi-cuadrado	9,375
gl	3
Sig. asintótica	,025

a. Coeficiente de concordancia de Kendall

Por último, en la tabla 109 se presentan las probabilidades obtenidas de la aplicación de la prueba de McNemar a cada pareja de variables. Tal y como se observa, la única pareja de variables que

muestra diferencias significativas es la que presenta valores extremos: en el extremo inferior, las respuestas del 2016 y, en el extremo superior, las respuestas del 2019 (p=,016). Salvo esta comparación, no se evidencian diferencias significativas.

Tabla 109. Resultados de las pruebas de McNemar para cada pareja de mediciones

V1	V2	*p*
a2016	a2017	0,125
a2016	a2018	0,375
a2016	a2019	0,016*
a2017	a2018	1,000
a2017	a2019	0,375
a2018	a2019	0,289

Se expresan los resultados

Para expresar los resultados de la prueba Q de Cochran puede usarse el siguiente formato:

Q(<gl>)= <Valor Q> p </= <valor p> W=<valor W>

Utilizando este formato, los anteriores resultados podrían ser descritos como sigue:

> *Para el examen de las diferencias entre el porcentaje de docentes que se expresan positivamente acerca de sus relaciones con los directivos docentes a través de cuatro años consecutivos, se aplicó una prueba Q de Cochran. Los resultados de la prueba mostraron diferencias globales significativas, con tamaños del efecto pequeños Q(3)=9,38 p=,025 W=,08. El porcentaje de docentes que se expresan positivamente sobre sus relaciones con los directivos inicia con un mínimo en el 2016 (52 %), aumenta en el 2017 (60 %), se mantiene para el 2018 en valores similares (60%) y para el 2019 alcanza un máximo de 70 %. Examinadas las pruebas de McNemar para el examen de diferencias entre estos años, solo se detectan diferencias significativas entre los años 2016 y 2019 (p=,016); para todos los otros casos las diferencias no son significativas.*

Capítulo 12

Análisis de varianza con más de una variable independiente

Hasta el momento, hemos recorrido un largo camino. En la primera parte del capítulo 10 se presentaron las pruebas que contrastan una variable dependiente en dos segmentos diferentes de población y en la primera parte del capítulo 11 se indicaron las extensiones de esas pruebas para el caso de tres o más segmentos de población. En ambos se han examinado los cambios de una variable dependiente entre dos o más grupos, definidos por un único factor independiente.

Ahora bien, en la segunda parte del capítulo 10 se presentaron las pruebas que contrastan dos aplicaciones "apareadas" de la misma variable dependiente en la misma población. En la segunda parte del capítulo 11 extendimos esto para considerar tres o más aplicaciones apareadas. Esto nos permitió examinar los diseños "intrasujeto".

En este capítulo se enseñarán algunas pruebas, relativamente simples, en las que combinamos todos los elementos anteriores mientras seguimos examinando diferencias en una misma variable dependiente.

Se presentarán únicamente tres pruebas de uso muy frecuente en la investigación educativa y social que, aunque involucran —o "controlan"— dos o más variables en su influencia sobre la variable dependiente, habitualmente no se incluyen en los manuales de estadística multivariante: el análisis factorial de varianza, el análisis mixto de varianza y el análisis de covarianza.

El análisis factorial de varianza, también llamado Anova de dos vías, o Anova factorial, es una prueba paramétrica que tiene una variable dependiente numérica y continua, y puede presentar dos o más variables independientes (factores) que se cruzan y generan diferentes grupos en su intersección. Es una prueba muy popular que permite el análisis de los llamados "diseños factoriales de investigación".

Por su parte, el análisis mixto de varianza, o Anova mixto, puede ser entendido como una extensión del Anova factorial en el que ahora también se consideran mediciones intrasujeto, del tipo pretest y postest, por ejemplo. Desde otro punto de vista, el Anova mixto podría también entenderse como una extensión del Anova de medidas repetidas, en el cual además hemos incluido

factores intersujeto. En los dos casos tenemos un factor intrasujeto (por ejemplo, pretest, una medición intermedia y postest) y uno, o varios, factores intersujeto (por ejemplo grupo experimental/control, masculino/femenino, etc.).

Es importante anotar que la denominación de Anova "mixto" puede ser confusa, ya que también puede representar un Anova en el que se combinan efectos "fijos" y efectos "aleatorios". Los efectos fijos son los que hemos venido tratando y se refieren a los diferentes tratamientos que se comparan. Los efectos aleatorios, por su parte, son una selección de una población de muchos y muy diferentes tratamientos con la que se establecen las comparaciones. En este contexto hablaremos de Anova mixto para referirnos al Anova que combina medidas intrasujetos e intersujetos.

Por último, hablaremos de una nueva extensión del Anova, conocida como análisis de covarianza, o Ancova. En esta se "controlan" los efectos de una nueva variable, métrica y continua, sobre la variable dependiente. Esta nueva variable se conoce como "covariable" y es la que da su nombre a este tipo de procedimiento. Al sustraer el efecto de la covariable sobre la variable dependiente podemos correr un Anova mixto sin esa interferencia, lo que nos permite aumentar la precisión de los experimentos.

Estos tres procedimientos son de uso muy común en la investigación educativa y social, pero no siempre son adecuadamente utilizados. En particular, es importante advertir sobre una confusión, bastante extendida, presente entre el Anova mixto y el Ancova. Muchos investigadores utilizan el Ancova para controlar el efecto del pretest sobre el postest, definiendo como variable dependiente las puntuaciones de postest y como covariable las puntuaciones del pretest. Esto, en sí, no es incorrecto, mientras los participantes hayan sido asignados aleatoriamente a los grupos. Sin embargo, cuando los sujetos no son asignados al azar a grupos, usar la covariación de la prueba previa no es apropiado porque las diferencias entre los grupos se producen no solo por variación de probabilidad, sino también debido a diferencias sistemáticas entre los grupos (Mara *et al.*, 2012).

Análisis factorial de varianza

Presentación

Diseños factoriales de investigación

El análisis factorial de varianza, Anova factorial, o Anova de dos vías, o dos factores, es la extensión del Anova de una vía, que examinamos en el capítulo 11 (Anova *one way*), pero ahora considerando, no un factor, sino dos o más, independientes entre sí, como variables independientes.

Para entender la importancia y las características del Anova factorial, deben comprenderse las características de los diseños factoriales de investigación. En un *diseño factorial de investigación* se analizan en el mismo momento, los efectos simultáneos de dos o más variables categóricas (factores) sobre una o varias variables dependientes. Para examinar estos efectos, en el diseño se forman grupos para cada combinación de valores de todos los factores considerados.

Así, por ejemplo, si se quisieran conocer los efectos de dos factores: sexo, con dos valores (masculino y femenino), y nivel educativo, con tres valores (secundaria, pregrado y posgrado), sobre el resultado de una prueba de ansiedad evaluativa, necesitaríamos que la muestra de estudio se distribuyera en todas las casillas de la matriz de la figura 69.

		Sexo	
		Masculino	**Femenino**
Nivel educativo	Secundaria		
	Pregrado		
	Posgrado		

Figura 69. Diseño factorial 2x3

En este caso, debemos tener un número suficiente de hombres y mujeres de cada uno de los tres niveles educativos. Este puede ser denominado como un "diseño 2x3 con una variable dependiente". Como se entiende, la denominación "2x3" hace referencia al número de valores de cada factor e indica también el número total de combinaciones, 6 casillas.

Evidentemente, podemos extender este diseño de investigación a una situación de tres factores independientes. Para ello, incluyamos en el diseño anterior un nuevo factor con, digamos, dos valores: la presencia de calificaciones relativamente altas o bajas, por ejemplo. Al incluir este nuevo factor, tenemos ahora un diseño factorial 2x2x3, con un total de 12 casillas. La representación gráfica puede ser como se muestra en la figura 70.

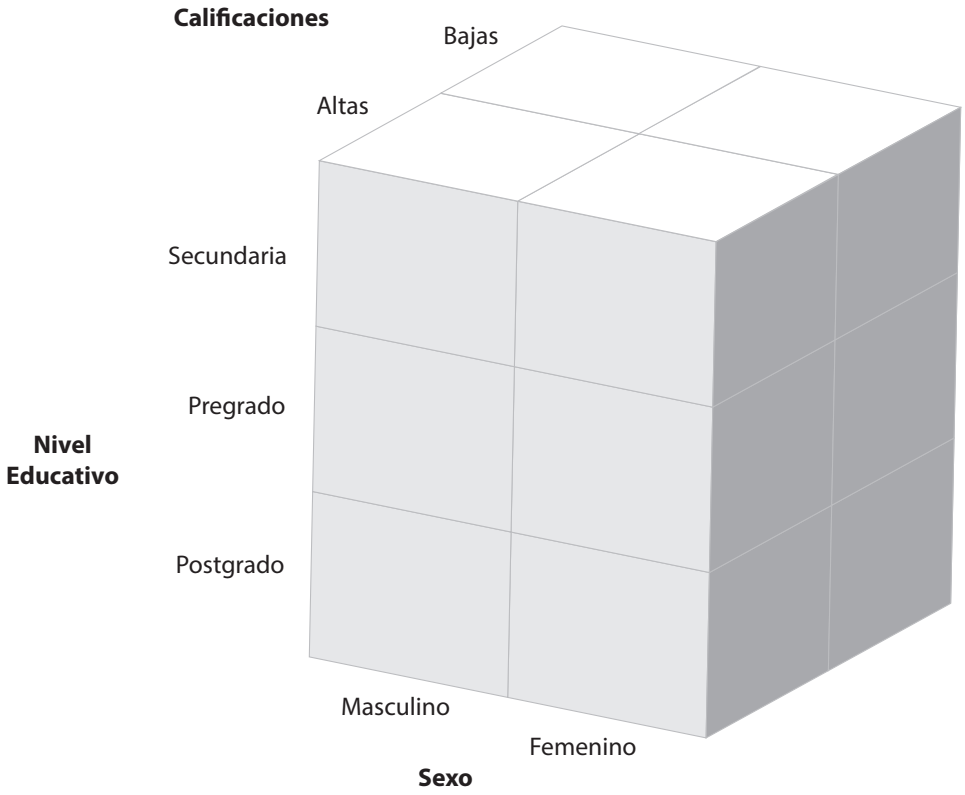


Figura 70. Diseño factorial 2x2x3

Aunque no sería tan clara la representación gráfica, no hay impedimentos para la consideración de diseños factoriales con más de tres factores, cuatro, cinco o más, aunque estos no son, en la investigación educativa y social, para nada frecuentes.

Por otro lado, es relevante tener en cuenta que en un diseño factorial de investigación no es estrictamente requerido que se considere una única variable dependiente. Por el contrario, podrían tomarse varias variables dependientes, si bien en estos casos daríamos, en el ámbito de la estadística, un salto considerable hacia los modelos de análisis multivariante de varianza (Manova), que presentan una complejidad considerablemente mayor que los modelos Anova que trataremos en este capítulo. Se espera hacer la presentación de este tipo de modelos multivariantes en el futuro, en un segundo tomo de la presente obra.

Las ventajas de los diseños factoriales de investigación sobre los diseños simples con un solo factor son evidentes. Primero, es clara la eficiencia del diseño factorial, en tanto no necesitamos considerar dos muestras separadas para el examen de dos factores independientes: con una sola muestra podemos examinar los efectos de los dos factores.

Segundo, existe una clara ventaja de los diseños factoriales sobre los diseños de un factor, derivada de la posibilidad de examinar, no solo los efectos separados de cada factor, sino los efectos de la combinación de los factores. Esto es bastante más de lo que lograríamos con la aplicación de dos Anova de una vía a nuestra muestra.

Para comprender los efectos de la combinación de factores debe considerarse, por un lado, un diseño factorial en el que cada uno de los dos factores tuviera un efecto apreciable sobre la variable dependiente. Es posible también que su combinación no aporte nada nuevo: en ese caso el efecto de la combinación sería, simplemente, la suma de los efectos separados de cada variable. Es posible, sin embargo, que la combinación de las variables cambie el resultado: lo intensifique, por ejemplo, o lo neutralice en alguna medida. Esto es, precisamente, lo que podemos detectar en un diseño factorial de investigación: el efecto de la interacción de los factores.

Explicaremos esto en el diseño 2x3 que planteamos originalmente, en el que proponemos el examen del efecto del sexo, con dos valores, y el nivel educativo, con tres, sobre la ansiedad educativa. Primero, el examen del efecto del sexo sobre la ansiedad evaluativa indica que las mujeres parecen mostrar mayores niveles de ansiedad que los varones. Segundo, el examen del efecto del nivel educativo muestra que en el nivel universitario de pregrado se encuentran niveles de ansiedad mucho mayores a los observados en secundaria o posgrado. Finalmente, el examen conjunto de las dos variables indica que la diferencia entre varones y mujeres se acentúa de forma muy importante en secundaria, y va bajando en los niveles superiores, hasta llegar al mínimo en el nivel de posgrado. Tenemos, en este caso, que cada factor tiene efectos significativos, y que la interacción de los factores también lo tiene. Examinaremos este caso más adelante en el ejemplo.

Un punto adicional para considerar. Cuando tenemos dos factores, el modelo examinará el efecto de cada factor (dos efectos) y el de su interacción (un efecto); en total tres efectos. Cuando tenemos tres factores, los exámenes se multiplican, por cuanto deberán examinarse los efectos de cada factor (tres efectos), los de cada pareja de factores (tres efectos: F1/F2, F1/F3 y F2/F3) y los de la interacción de los tres factores (un efecto: F1/F2/F3), en total siete efectos. Ya para cuatro factores, deberán examinarse quince efectos.

El Anova factorial

El *análisis factorial de varianza* o *Anova factorial* es una prueba de hipótesis paramétrica que examina diferencias en las medias de una variable numérica continua respecto de dos o más factores categoriales y sus interacciones. Dado que el Anova factorial es una prueba paramétrica, requiere del cumplimiento de los *supuestos* propios de este tipo de pruebas. En particular, tiene los siguientes supuestos:

- Una variable dependiente, numérica y continua.
- Dos, o más, variables independientes categoriales (factores), que definen un total de *m* poblaciones diferentes, donde *m* es resultado de la multiplicación del número de valores de los distintos factores independientes.
- El modelo supone que las observaciones han sido aleatoriamente seleccionadas y, por tanto, son independientes entre sí.
- La variable dependiente se distribuye de forma aproximadamente normal en las *m* poblaciones.
- La varianza de la variable dependiente es igual entre las *m* poblaciones (homocedasticidad).

Respecto del supuesto de normalidad, debe anotarse que, aunque el Anova factorial es una prueba relativamente robusta frente a la violación de este supuesto, sí se requiere, al menos, que la distribución sea simétrica y no presente valores atípicos muy significativos en cada una de las poblaciones consideradas. Cumplida esta condición, puede correrse la prueba. Si definitivamente esto no es posible, deberemos aplicar transformaciones a la variable. Lamentablemente, no existe una prueba no paramétrica equivalente al Anova factorial.

En cuanto al cumplimiento del supuesto de homocedasticidad no existen, en este caso, correcciones a la prueba como sí las encontrábamos en el Anova de una vía. En el caso en el que no se pueda verificar este supuesto, deberemos bien proceder a la ejecución de transformaciones, o bien a la consideración de que ese análisis de varianza presentará un incremento importante de las posibilidades de cometer un error de tipo I; esto es, que se han incrementado las posibilidades de rechazar la hipótesis nula siendo esta verdadera.

En el Anova factorial se ponen a prueba varias hipótesis nulas: una para cada uno de los factores involucrados y una para cada una de las interacciones. Esto hace que en un Anova factorial de dos factores tengamos tres hipótesis nulas y tres alternativas; en un Anova de tres factores tendremos siete hipótesis nulas y sus correspondientes alternativas. En cada caso, la hipótesis nula de cada factor afirmará que no hay diferencias en las medias entre los grupos, mientras que la alternativa dirá que, al menos uno de los grupos, muestra diferencias significativas con los otros. Por su parte, la hipótesis nula de la interacción afirmará que no hay ningún efecto de interacción significativo.

Al correr un Anova factorial, el *software* estadístico aportará información separada para cada uno de los factores y para cada una de sus interacciones. En cada caso se presenta el valor de estadístico (*F*), sus grados de libertad (gl), el nivel de probabilidad asociado en la distribución (*p*) y una o varias medidas de tamaño del efecto.

Como en el caso de los otros análisis que hemos examinado para más de dos medias, el Anova factorial nos afirmará que hay diferencias entre los grupos, pero no indicará entre qué grupos específicos se verifican diferencias significativas. Para saberlo deben ser examinadas pruebas *post hoc*.

Como en los casos anteriores y a riesgo de ser considerados excesivamente reiterativos, las pruebas *post hoc* solo pueden ser examinadas en los casos en que el Anova factorial haya indicado la presencia de diferencias significativas ligadas a un factor o a una interacción específica. Ahora, si se verifican diferencias globales, y en el factor solo existen dos valores como máximo, no se requiere la aplicación de pruebas *post hoc*, por cuanto la diferencia global se da entre los dos valores. Si, por el contrario, se comprueban diferencias globales y el factor tiene tres o más valores, las pruebas *post hoc* especificaran las parejas de valores específicos en donde se presentan diferencias significativas.

Un punto adicional sobre la aplicación de las pruebas *post hoc* es el relacionado con las pruebas específicas que deben ser aplicadas. Si, para el Anova, se verificó el cumplimiento del supuesto de homocedasticidad, podemos proceder con las pruebas *post hoc* estándar, o habituales; esto es, las pruebas con corrección de Tukey, Scheffe, Bonferroni, Holm o Sidak. Se sugiere, en este caso, el uso de la prueba de Tukey. Si, por el contrario, el supuesto de homocedasticidad no pudo ser verificado, las pruebas *post hoc* por elegir cambian; las alternativas son ahora las pruebas de Games-Howell, Dunnet o Dunn. Se sugiere en este último caso el uso de la prueba de Games-Howell.

Respecto de las medidas de tamaño de efecto, estas deben ser reportadas con cada hipótesis. Se conservan las recomendaciones dadas en los capítulos anteriores. Específicamente, y dependiendo del programa con el que se procesen las pruebas, existen diferentes alternativas para las medidas de tamaño del efecto. El programa JASP permite tres opciones para el cálculo del tamaño del efecto en un Anova factorial: el eta cuadrado (η^2), el eta parcial al cuadrado (η_p^2) y el omega cuadrado (ω^2). En el IBM-SPSS solo existe la alternativa del eta cuadrado parcial.

El η^2 es una medida bastante popular que resulta de fácil interpretación porque expresa el porcentaje de varianza explicada (R^2), pero con su uso se dificulta la comparación del efecto de la misma variable en distintos estudios. Por esta razón, es preferible el uso del eta al cuadrado parcial (η_p^2) que resuelve este problema, aunque no es tan fácil de interpretar. Sin embargo, cuando las muestras son de tamaño pequeño ($n<30$), η_p^2 tiende a sesgarse, por lo que, en estos casos, se prefiere el uso del ω^2. La tabla 110 permite interpretar estas medidas de tamaño del efecto en el Anova factorial. Como se observa, los límites de interpretación del η_p^2 y del ω^2 son idénticos.

Tabla 110. Límites para la interpretación de las medidas de tamaño del efecto en el Anova de una vía

Medida de tamaño del efecto	Nulo	Pequeño	Mediano	Grande
η^2	<0,1	0,1	0,25	0,37
η_p^2 (n>30)	<0,01	0,01	0,06	0,14
ω^2 (n<30)	<0,01	0,01	0,06	0,14

Finalmente, algunas palabras sobre las gráficas usadas para mostrar diferencias. Tradicionalmente, se han utilizado gráficos de líneas para trazar las medias de los distintos grupos en el análisis. Este tipo de gráficas es el que aparece, por defecto, tanto en el JASP como en el IBM-SPSS. Sin embargo, actualmente se recomiendan las gráficas de barras para este mismo efecto, en la medida en que las líneas sugieren una continuidad que la que carecen los factores (Aron y Aron, 2001). En la medida en que estos gráficos rara vez se publican, y con frecuencia solo se utilizan para ayudar al investigador a interpretar los resultados, se sugiere el uso de la gráfica que resulte más cómoda para tal efecto.

Cómo ejecutar un Anova factorial

Para examinar un análisis de factorial de varianza en los diferentes programas puede procederse de la forma presentada en el recuadro 47, para el programa JASP. En el programa IBM-SPSS debe buscarse el procedimiento a través del menú "Modelo lineal general/Univariante…" (recuadro 48).

Recuadro 47. Cómo ejecutar un Anova factorial en JASP

/ANOVA/ANOVA...

En este punto debe pasarse la variable dependiente a la lista "dependent variable" y los factores independientes a la lista "fixed factors"

Display

- √ Descriptive statistics
- √ Estimates effect size
  - √ Partial η_p^2 o √ ω^2 (dependiendo del tamaño de la muestra)
- Assumption Checks
  - √ Homogeneity test
  - √ Q-Q plot of residuals
- Post Hoc test
  - (dependiendo de la homogeneidad, se selecciona la prueba adecuada)
  - √ Tukey o √ Games-Howell
- Descriptive Plots (pasar la variable de factor a "Horizontal Axis")
  - √ Display error bars

Recuadro 48. Cómo ejecutar un Anova factorial en IBM-SPSS

/Analizar/Modelo lineal general/Univariante…
En este punto debe pasarse la variable dependiente a la casilla "variable dependiente" y los diferentes factores a la lista "factores fijos"

- En el botón "Gráficos"
  Pasar de la lista "Factores" a un factor a "Eje horizontal" y otro a "Líneas separadas". Es posible pasar un tercer factor a "Gráficos separados"
  Pulsar "Continuar"
- En el botón "Post hoc"…
  Deben seleccionarse los factores sobre los que se desean la pruebas *post hoc* y seleccionar la prueba adecuada
  √ Tukey o √ Games-Howell
  Pulsar "Continuar"
- En el botón "Opciones":
  √ Estadísticos descriptivos
  √ Pruebas de homogeneidad
  √ Estimaciones del tamaño del efecto
  Pulsar "Continuar"

Pulsar "Aceptar"

El ejemplo: diferencias en el aprendizaje por sexo y nivel educativo

En el capítulo 9 se examinaron los resultados de un estudio real en que una muestra de estudiantes, diferenciados por el nivel educativo que se encontraban cursando (secundaria, pregrado y posgrado), que presentaron la prueba MSLQ. Esta prueba examina diferentes características del aprendizaje.

En este caso extenderemos el ejemplo anterior para considerar la influencia simultánea del nivel educativo, con tres valores, y del género, con dos, sobre seis diferentes características del aprendizaje, a saber:

- *El uso de metas intrínsecas*. Entendido como la tendencia que presenta el estudiante a aprender, por el interés mismo de hacerlo, sin motivación externa.
- *El uso de metas extrínsecas*. Entendido como con el interés que tiene el estudiante hacia la materia, relacionado con ganar prebendas o evitar castigos o situaciones indeseables.
- *El valor de la tarea.* Indica la importancia dada por el estudiante a los contenidos mismos de la asignatura.
- *La ansiedad evaluativa.* Revela la experiencia emocional del estudiante durante las evaluaciones.
- *Los niveles de organización.* Señala los niveles de organización de los materiales de estudio por el estudiante.
- *Aprendizaje en parejas*. Muestra la tendencia del estudiante a hacer sus labores estudiantiles en parejas de trabajo.

En total, se cuenta con información de 597 estudiantes, de los cuales 318 (53,3 %) son hombres y 279 (46,7 %) son mujeres, 270 (45,2 %) se encuentran cursando la secundaria, 248 (41,5 %) se encuentran en el pregrado universitario y 79 (13,2 %) adelantan estudios de posgrado.[11]

Tenemos así un diseño factorial 2x3 con seis variables dependientes, que serán examinadas por separado. La distribución de la muestra en el cruce de los dos factores aparece en la tabla 111.

Tabla 111. Muestra discriminada por el cruce entre nivel educativo y género

	Nivel educativo			
Género	Secundaria	Pregrado	Posgrado	Total
Hombre	139	138	41	318
Mujer	131	110	38	279
Total	270	248	79	597

Puede ser importante anotar que, en otras condiciones, podría ser deseable examinar, para esta situación un análisis multivariante de varianza (Manova) para, en un solo procedimiento, examinar todas las influencias sobre las seis variables dependientes. No se hará así en esta ocasión, con propósitos pedagógicos. El examen de seis variables dependientes permitirá analizar diferentes tipos de influencias.

Planteamiento de las hipótesis

Se plantearán las hipótesis relacionadas con la primera de las escalas del MSLQ: el uso de metas intrínsecas. Las hipótesis de las otras cinco escalas siguen idéntica estructura. Iniciando con las hipótesis relacionadas con el factor de género:

> *Hipótesis nula 1 (H_{0_1}). No existen diferencias en las medias de la escala de uso de metas intrínsecas entre los grupos de género.*
>
> *Hipótesis alternativa 1 (H_{a_1}). Existen diferencias entre las medias de la escala de uso de metas intrínsecas entre los grupos de género.*

Las hipótesis relacionadas con el nivel educativo son levemente diferentes, por la presencia de tres valores:

> *Hipótesis nula 2 (H_{0_2}). No existen diferencias en las medias de la escala de uso de metas intrínsecas entre los grupos de nivel educativo.*
>
> *Hipótesis alternativa 2 (H_{a_2}). La media de la escala de uso de metas intrínsecas de al menos uno de los grupos de nivel educativo es diferentes de los otros grupos.*

11 Tanto el proyecto de investigación como los datos concretos son reales y provienen de poblaciones de estudiantes de Bogotá, Colombia, y fueron previamente publicados. Los resultados originales pueden ser consultados en Hederich-Martínez *et al.* (2018).

Las hipótesis relacionadas con las diferencias ligadas a las interacciones de los dos factores podrían ser planteadas de forma general, como una hipótesis nula y una alternativa por interacción. Si se hicieran de manera detallada, para cada interacción se tendrían que plantear todas las hipótesis ligadas a todas las posibles combinaciones entre los grupos. Para este caso, en el que tenemos un diseño 2x3, esto nos generaría un total de quince hipótesis nulas y otras tantas alternativas.

Hipótesis nula 3 (H_{0_3}). No existen diferencias en las medias de la escala de uso de metas intrínsecas entre los grupos de cualquiera de los factores.

Hipótesis alternativa 3 (H_{0_3}). Existen diferencias en las medias de la escala de uso de metas intrínsecas entre dos grupos de cualquiera de los factores.

Selección de la prueba

Se examinarán los supuestos propios del Anova factorial de las seis pruebas planteadas de forma paralela. Como se recuerda, el Anova factorial requiere que la variable dependiente se distribuya de forma aproximadamente normal, sin valores atípicos muy significativos, y presente homocedasticidad entre los grupos definidos por el diseño.

Iniciando con el supuesto de normalidad aproximada, en la figura 71 se observan los gráficos Q-Q de cada una de las seis variables dependientes.

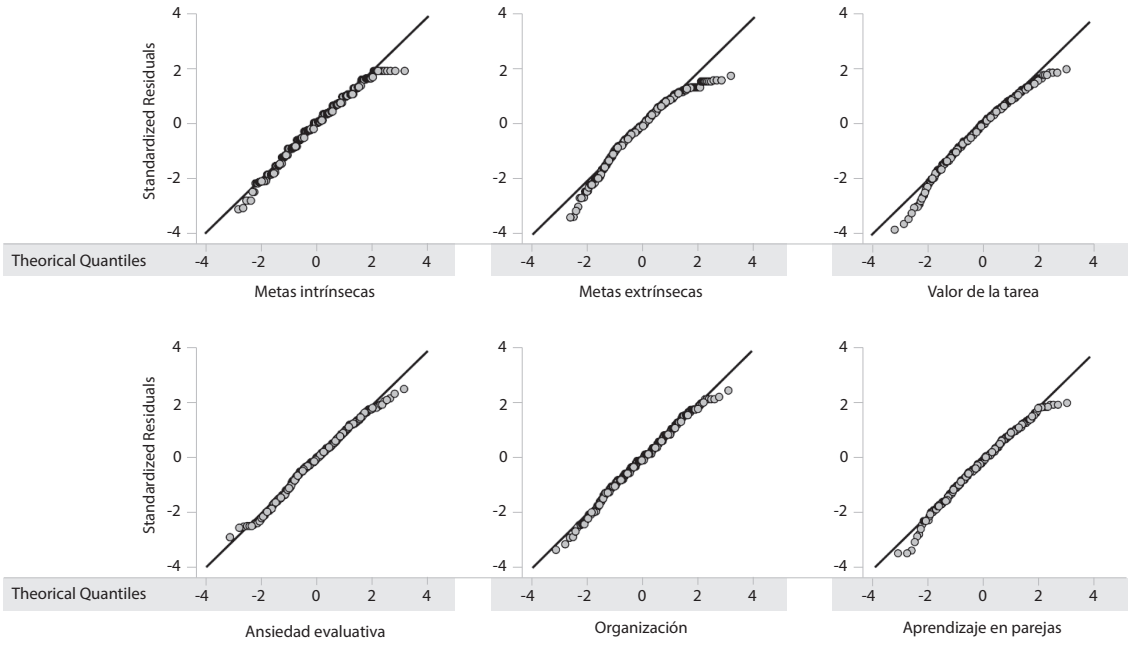


Figura 71. Gráficas Q-Q de seis variables dependientes

Se observan leves desviaciones a la normalidad en las gráficas de metas intrínsecas, metas extrínsecas y aprendizaje en parejas con muy pocos casos en el límite superior que se separan levemente de la diagonal. Puede asumirse que se cumple el supuesto.

El segundo supuesto es el de la homocedasticidad de las variables dependientes entre los diferentes grupos del diseño. Para examinar este supuesto, se corren pruebas de Levene para cada variable dependiente. La tabla 112 presenta los resultados de las seis pruebas.

Tabla 112. Resultados de las seis pruebas de Levene para la homocedasticidad de las variables dependientes

Variable	F	gl1	gl2	p
Metas intrínsecas	1,62	5	591	0,151
Metas extrínsecas	1,56	5	591	0,169
Valor de la tarea	3,01	5	591	0,011*
Ansiedad evaluativa	2,90	5	591	0,013*
Organización	2,98	5	591	0,011*
Aprendizaje en parejas	1,81	5	591	0,109

Como se observa, de las seis pruebas examinadas, tres cumplen con el supuesto de homocedasticidad y tres no lo cumplen: las escalas de valor de la tarea, ansiedad evaluativa y uso de estrategias de organización. Esto deberá ser tenido en cuenta en el momento de examinar pruebas *post hoc* para estas variables.

Las gráficas de la figura 72 muestran las diferencias entre las medias para cada grupo. Los tres niveles educativos están representados en el eje x, mientras que los dos géneros aparecen representados por líneas diferentes en cada gráfica.

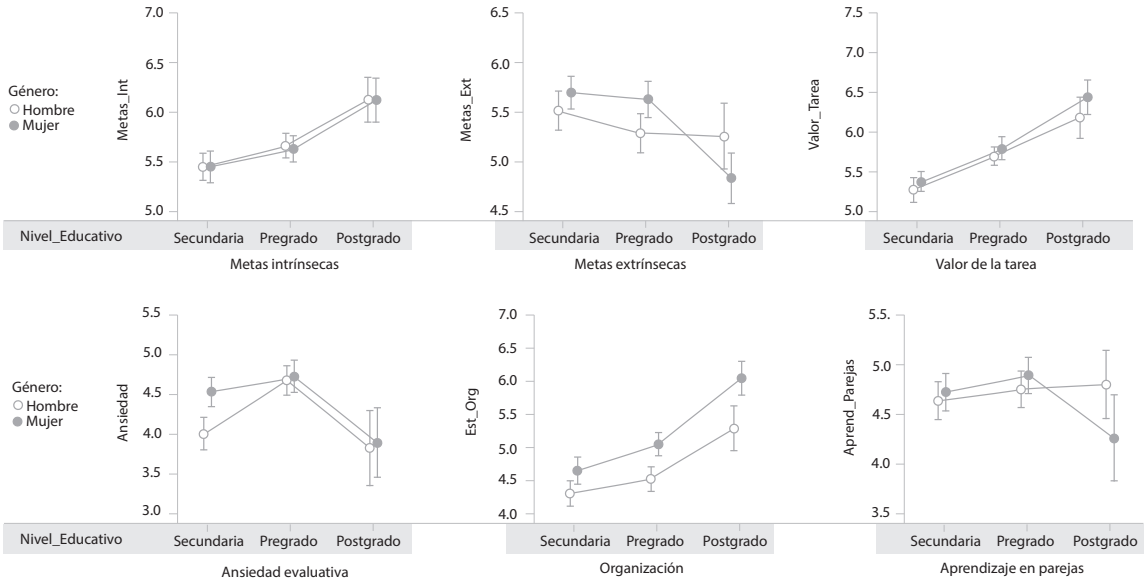


Figura 72. Medias de las variables dependientes por sexo y nivel educativo

Una inspección rápida de las gráficas anteriores muestra comportamientos disímiles entre las diferentes escalas. Se observan comportamientos estrictamente crecientes en cuanto a nivel educativo

para las escalas de metas intrínsecas, valor de la tarea y organización y comportamientos levemente decrecientes en metas extrínsecas. De igual forma, parece mostrarse comportamientos no lineales con un máximo en el nivel de pregrado para las escalas de ansiedad evaluativa y trabajo en parejas. En cuanto al género, tienden a marcarse mayores distancias entre los grupos de género en metas extrínsecas y organización.

Las posibles diferencias en las interacciones entre los dos factores podrían ser sospechadas cuando en dos grupos específicos parecen mostrarse diferencias que están ausentes en los otros. En ese sentido, podríamos encontrar diferencias ligadas a la interacción en las escalas de metas extrínsecas, ansiedad evaluativa y aprendizaje en parejas.

La tabla 113 muestra los estadísticos descriptivos de las seis escalas para los diferentes grupos presentes en el diseño. Estos datos serán relevantes en el momento de reportar los resultados.

Tabla 113. Estadísticos descriptivos para las seis escalas en cada grupo

Género	N. educativo	Metas intrínsecas		Metas extrínsecas		Valor de la tarea	
		Media	DE	Media	DE	Media	DE
Hombre	Posgrado	6,12	0,71	5,25	1,04	6,18	0,82
	Pregrado	5,65	0,74	5,28	1,16	5,69	0,67
	Secundaria	5,44	0,81	5,51	1,16	5,27	0,92
Mujer	Posgrado	6,11	0,68	4,83	0,78	6,43	0,65
	Pregrado	5,62	0,71	5,62	0,94	5,80	0,75
	Secundaria	5,44	0,90	5,69	0,96	5,37	0,74

Género	N. educativo	Ansiedad		Organización		Ap. parejas	
		Media	DE	Media	DE	Media	DE
Hombre	Posgrado	3,83	1,49	5,28	1,06	4,80	1,08
	Pregrado	4,67	1,07	4,52	1,05	4,75	1,10
	Secundaria	4,00	1,22	4,30	1,13	4,64	1,11
Mujer	Posgrado	3,90	1,34	6,04	0,76	4,26	1,33
	Pregrado	4,73	1,07	5,05	0,91	4,90	0,95
	Secundaria	4,53	1,06	4,65	1,16	4,73	1,07

Cálculo de resultados

Se han reunido los resultados generales de todas las pruebas Anova factorial en la tabla 114.

Tabla 114. Tablas Anova factorial para el examen de diferencias ligadas a sexo y nivel educativo en los puntajes de metas intrínsecas, metas extrínsecas, valor de la tarea, ansiedad, organización y trabajo en parejas

Escala	Fuente	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
Metas intrínsecas	Género	0,017	1	0,017	0,027	0,869	4,607e -5
	NE	**28,381**	**2**	**14,191**	**22,761**	**< ,001**	**0,072**
	Ge*NE	0,036	2	0,018	0,029	0,971	9,867e -5
	Residuos	368,471	591	0,623			
Metas extrínsecas	Género	0,123	1	0,123	0,111	0,739	1,873e -4
	NE	**19,196**	**2**	**9,598**	**8,636**	**< ,001**	**0,028**
	Ge*NE	**8,630**	**2**	**4,315**	**3,882**	**0,021**	**0,013**
	Residuos	656,855	591	1,111			
Valor de la tarea	**Género**	**2,562**	**1**	**2,562**	**4,222**	**0,040**	**0,007**
	NE	**65,174**	**2**	**32,587**	**53,716**	**< ,001**	**0,154**
	Ge*NE	0,412	2	0,206	0,339	0,712	0,001
	Residuos	358,531	591	0,607			
Ansiedad	**Género**	**5,093**	**1**	**5,093**	**3,796**	**0,052**	**0,006**
	NE	**49,515**	**2**	**24,757**	**18,451**	**< ,001**	**0,059**
	Ge*NE	**8,176**	**2**	**4,088**	**3,047**	**0,048**	**0,010**
	Residuos	793,013	591	1,342			
Organización	**Género**	**32,292**	**1**	**32,292**	**28,621**	**< ,001**	**0,046**
	NE	**86,137**	**2**	**43,069**	**38,172**	**< ,001**	**0,114**
	Ge*NE	2,952	2	1,476	1,308	0,271	0,004
	Residuos	666,807	591	1,128			
Aprendizaje en parejas	Género	1,177	1	1,177	0,995	0,319	0,002
	NE	5,866	2	2,933	2,478	0,085	0,008
	Ge*NE	**7,447**	**2**	**3,723**	**3,146**	**0,044**	**0,011**
	Residuos	699,364	591	1,183			

NE: nivel educativo.
Los efectos significativos (p<,052) aparecen en negrilla y sombreados.

Como se observa, tenemos casi todas las posibilidades de combinación en esas seis pruebas. La escala de metas intrínsecas muestra efectos significativos del nivel educativo. La escala de metas extrínsecas evidencia efectos significativos del nivel educativo y de su interacción con el género. Las escalas de valor de la tarea y organización revelan diferencias significativas en género y nivel educativo, pero no en su interacción. La escala de ansiedad expone todos los efectos significativos: género (a *p*=,052), nivel educativo y su interacción. Finalmente, la escala de aprendizaje por parejas solo expresa efectos significativos en la interacción de los dos factores principales.

Iniciando por la escala de metas intrínsecas, solo se observan efectos significativos asociados con el nivel educativo. En este sentido, la media de la escala es creciente: a medida en que se avanza en el nivel educativo aumenta la media de la escala de metas intrínsecas. Ni el género ni su interacción con el nivel educativo muestran influencias significativas.

Dado que la variable superó el supuesto de homocedasticidad, es posible calcular pruebas *post hoc* estándar con corrección de Tukey sobre los niveles educativos. El resultado de estas pruebas se presenta en la tabla 115.

Tabla 115. Pruebas post hoc para el examen de diferencias entre los niveles educativos

Comparaciones *post hoc*–Nivel_Educativo						
		gl	**SE**	**t**	**d de Cohen**	**p_{tukey}**
Secundaria	Pregrado	-0.198	0.070	-2.835	-0.247	0.013 *
	Posgrado	-0.679	0.101	-6.714	-0.823	< .001 ***
Pregrado	Posgrado	-0.481	0.102	-4.705	-0.665	< .001 ***

* p < .05, ** p < .01, *** p < .001

Notas: la *d* de Cohen no corrige comparaciones múltiples; el valor de *p* fue ajustado para comparar a una familia de 3; los resultados son en promedio superiores a los de "género".

Como se observa, a medida que se avanza el nivel educativo el uso de metas intrínsecas es mayor. Se presentan diferencias significativas entre secundaria y pregrado (*p*=,013), secundaria y posgrado (*p*<,001) y pregrado y posgrado (*p*<,001).

En segundo lugar, examinamos la escala de metas extrínsecas. Para este caso, se evidencian efectos significativos del nivel educativo y de su interacción con el género, pero no se presentan efectos significativos con el género, como efecto principal. El examen de la gráfica muestra que la línea de los hombres parece ser levemente descendente entre los niveles educativos, mientras que la línea de las mujeres, que se sitúa inicialmente por encima de la de los hombres entre secundaria y pregrado, presenta una brusca caída en el nivel de posgrado.

Para el examen de las pruebas *post hoc* indicadas, debe recordarse que la escala de metas extrínsecas superó el supuesto de homocedasticidad, por lo que podemos utilizar pruebas *post hoc* estándar con corrección de Tukey. Los resultados se presentan en las tablas 115 y 116.

Tabla 116. Resultados de las pruebas post hoc para el examen de diferencias entre los niveles educativos

Comparaciones *post hoc*–Nivel_Educativo						
		DE	SE	t	d de Cohen	p_{tukey}
Secundaria	Pregrado	0.150	0.093	1.610	0.139	0.242
	Posgrado	0.560	0.135	4.147	0.535	< .001 ***
Pregrado	Posgrado	0.410	0.136	3.003	0.390	0.008 **

* p < .05, ** p < .01, *** p < .001

Notas: la *d* de Cohen no corrige comparaciones múltiples; el valor de *p* fue ajustado para comparar a una familia de 3; los resultados son en promedio superiores a los de "género".

Tabla 117. Resultados de las pruebas **post hoc** *para el examen de diferencias entre los niveles educativos según el género*

Comparaciones *post hoc*–Género * Nivel_Educativo					
		DE	SE	t	p_{tukey}
Hombre, secundaria	Mujer, secundaria	-0.182	0.128	-1.419	0.715
	Hombre, pregrado	0.228	0.127	1.801	0.466
	Mujer, pregrado	-0.111	0.135	-0.822	0.963
	Hombre, posgrado	0.258	0.187	1.379	0.740
	Mujer, posgrado	0.679	0.193	3.518	0.006 **
Mujer, secundaria	Hombre, pregrado	0.410	0.129	3.191	0.019 *
	Mujer, pregrado	0.072	0.136	0.525	0.995
	Hombre, posgrado	0.440	0.189	2.335	0.182
	Mujer, posgrado	0.861	0.194	4.433	< .001 ***
Hombre, pregrado	Mujer, pregrado	-0.339	0.135	-2.514	0.122
	Hombre, posgrado	0.030	0.188	0.161	1.000
	Mujer, posgrado	0.451	0.193	2.334	0.182
Mujer, pregrado	Hombre, posgrado	0.369	0.193	1.912	0.396
	Mujer, posgrado	0.789	0.198	3.980	0.001 **
Hombre, posgrado	Mujer, posgrado	0.421	0.237	1.772	0.485

* p < .05, ** p < .01, *** p < .001

Notas: el valor de *p* fue ajustado para comparar a una familia de 6.

Como se observa en las tablas, sobre el efecto del nivel educativo se presentan diferencias significativas entre los niveles de secundaria y pregrado con el nivel de posgrado, en el sentido en que este último revela menores medias en la escala de metas extrínsecas (*p*<,001 y *p*=,008, respectivamente). El examen de las interacciones significativas muestra, por su parte, que las mujeres del nivel de posgrado presentan medias más bajas en la escala de metas extrínsecas, lo cual indica diferencias significativas con las mujeres en pregrado (*p*<,001) y con mujeres y hombres en secundaria

(*p*<,001 y *p*=,006, respectivamente). Además, se observa una diferencia significativa entre las mujeres de secundaria y los hombres de pregrado en el sentido en que las primeras muestran medias de metas extrínsecas mayores que los segundos.

Continuando con la escala de valor de la tarea, los resultados mostraron efectos principales significativos ligados al género y al nivel educativo, sin que se observen efectos significativos ligados a la interacción de estos dos factores. En relación con el género, las mujeres parecen asignar mayor valor a la tarea que sus compañeros varones en todos los niveles. Por parte del nivel educativo, los resultados muestran un comportamiento estrictamente creciente, que inicia con los niveles más bajos en secundaria, sigue con niveles intermedios en pregrado y culmina con los niveles máximos en posgrado.

Para el caso del valor de la tarea, la prueba de Levene para esta escala mostró que debía rechazarse la hipótesis nula de homocedasticidad, lo que lleva a examinar pruebas *post hoc* de Games-Howell. Los resultados indicaron la presencia de diferencias significativas entre todos los niveles educativos (*p*<,001 en todas las comparaciones). Las mayores diferencias se presentan entre los niveles de secundaria y posgrado.

En cuanto a la escala de ansiedad evaluativa, los resultados indicaron que todos los efectos principales muestran ser significativos, así como la interacción entre estos. Es importante anotar que aunque el nivel de significación ligado al género supera levemente el nivel convencional (p=,052), su proximidad con este permite aceptar esta diferencia. Al respecto, debe anotarse que las mujeres muestran mayores niveles de ansiedad, comparadas con los hombres en todos los niveles. Por su parte, se evidencian diferencias significativas entre los niveles educativos si bien el efecto no muestra ser lineal. Se inicia con niveles medios de ansiedad en secundaria, bastante más altos en mujeres, que se incrementan aún más en el pregrado. Para el posgrado los niveles de ansiedad muestran un pronunciado descenso, tanto en hombres como en mujeres.

El examen de las pruebas *post hoc* de Games-Howell muestra diferencias significativas entre secundaria y pregrado, así como entre pregrado y posgrado. El programa JASP no aporta pruebas *post hoc* de Games-Howell para el examen de las interacciones significativas. Aparentemente, los niveles de ansiedad de las mujeres, que muestran ser bastante más altos que los de los varones es secundaria, van haciéndose similares en los siguientes niveles educativos (tabla 118).

Tabla 118. Pruebas* post hoc *de Games Howell para el examen de diferencias entre los niveles educativos en Ansiedad

Comparaciones *post hoc* de Games Howell–Nivel_Educativo					
Comparación	**DE**	**SE**	**t**	**gl**	**p_{tukey}**
Secundaria–pregrado	-0.437	0.099	-4.426	515.990	< .001 ***
Secundaria–posgrado	0.399	0.175	2.283	111.246	0.062
Pregrado–posgrado	0.836	0.173	4.823	108.044	< .001 ***

*** p < .001

El caso de los puntajes en la escala de organización es similar al de la escala de valor de la tarea, ya que se observan todos los efectos principales significativos, sin que se presente una interacción significativa entre estos. En cuanto al género, las mujeres muestran mucho mayores niveles de organización que los hombres en todos los niveles educativos. El examen del nivel educativo, por su parte, muestra que, para los dos géneros, los niveles de organización van creciendo a medida que se avanza en el nivel. La tabla 119 presenta las pruebas *post hoc* de Games-Howell para esta variable en nivel educativo, que resultan ser las apropiadas en este caso por el no cumplimiento del supuesto de homocedasticidad.

Tabla 119. Pruebas post hoc de games Howell para el examen de diferencias entre los niveles educativos en Organización

Comparaciones *post hoc* de Games Howell–Nivel_Educativo					
Comparación	**DE**	**SE**	**t**	**gl**	**p_{tukey}**
Secundaria–pregrado	-0.285	0.096	-2.961	515.407	0.009 **
Secundaria–posgrado	-1.177	0.133	-8.837	144.562	< .001 ***
Pregrado–posgrado	-0.892	0.130	-6.835	134.368	< .001 ***

** p < .01, *** p < .001

Por último, debemos examinar la escala de aprendizaje en parejas. El examen de los efectos en esta escala muestra ser particularmente interesante, pues no se verifican efectos principales significativos, pero sí se observa que la interacción entre los dos efectos es significativa.

La tabla 120 muestra las pruebas *post hoc* estándar, con corrección de Tukey, que resultan apropiadas para el caso de la escala de aprendizaje en parejas, en la cual se verificó el supuesto de homocedasticidad. Según la información presentada, la única pareja de grupos que muestra diferencias significativas es la de las mujeres que se diferencian por su nivel educativo: pregrado y posgrado. Específicamente, se observa que las mujeres en el nivel de pregrado muestran una mayor tendencia a trabajar en parejas que las del nivel de posgrado.

Tabla 120. Pruebas post hoc estándar con corrección de Tukey para la escala de aprendizaje en parejas

Comparaciones *post hoc*–Género * Nivel_Educativo					
		DE	SE	t	p_{tukey}
Hombre, secundaria	Mujer, secundaria	-0.088	0.132	-0.661	0.986
	Hombre, pregrado	-0.114	0.131	-0.869	0.954
	Mujer, pregrado	-0.257	0.139	-1.851	0.434
	Hombre, posgrado	-0.162	0.193	-0.839	0.960
	Mujer, posgrado	0.379	0.199	1.904	0.401
Mujer, secundaria	Hombre, pregrado	-0.026	0.133	-0.196	1.000
	Mujer, pregrado	-0.169	0.141	-1.204	0.835
	Hombre, posgrado	-0.075	0.195	-0.384	0.999
	Mujer, posgrado	0.467	0.200	2.328	0.184
Hombre, pregrado	Mujer, pregrado	-0.143	0.139	-1.031	0.907
	Hombre, posgrado	-0.049	0.193	-0.251	1.000
	Mujer, posgrado	0.493	0.199	2.472	0.134
Mujer, pregrado	Hombre, posgrado	0.095	0.199	0.476	0.997
	Mujer, posgrado	0.636	0.205	3.107	0.024
Hombre, postgrado	Mujer, posgrado	0.541	0.245	2.210	0.235

* $p < .05$

Notas: el valor de *p* fue ajustado para comparar a una familia de 6.

Interpretación y reporte

Para la presentación de los resultados de un Anova factorial, se escribe el texto y su interpretación y, al final de este, se anexan los resultados de la prueba con el siguiente formato:

$$F(gl1, gl2) = <\text{Valor F}> \; p <= <\text{Valor } p> \; \eta_p^2 = <\text{Valor} > \; \eta_p^2$$

Para el caso de los primeros dos ejemplos que se han desarrollado, la expresión de los resultados en texto podría quedar de la siguiente forma:

> *Se examinó un Anova factorial manteniendo como variable dependiente la escala de metas intrínsecas. Los resultados indicaron diferencias significativas asociadas con el nivel educativo que revelan un tamaño del efecto entre mediano y grande F(2,591)=14,19 p<,001* η_p^2 *=,07. La prueba de Levene demostró la posibilidad de examinar pruebas post hoc estándar con corrección de Tukey F(5,591)=1,56 p=,151. Los resultados de las pruebas* post hoc *señalaron que la media de la escala muestra un comportamiento estrictamente creciente con diferencias significativas entre todos los niveles (p=,013 entre secundaría y pregrado y p<,001 entre secundaria y posgrado, así como entre pregrado y posgrado). No se verifican diferencias significativas ligadas al género F(1,591)=0,03 p=,869* η_p^2 *<,01, ni a la interacción entre género y nivel educativo F(2,591)=0,03 p=,971* η_p^2 *<,01.*
>
> *En la escala de metas extrínsecas los resultados mostraron diferencias significativas ligadas al nivel educativo, con tamaños del efecto entre pequeños y medianos F(2,591)=8,64 p<,001*

η_p^2 =,03 y de su interacción con el género, con tamaños del efecto pequeños F(2,591)=3,88 p=,021 η_p^2 =,01. No se presentan diferencias significativas ligadas al género como efecto principal F(1,591)=0,11 p=,739 η_p^2 <,01.

Las medias de la escala de metas extrínsecas muestran ser similares entre secundaria y pregrado. Para el posgrado, la media de los hombres se mantiene en niveles similares al pregrado mientras que la de las mujeres presenta una brusca caída. La escala de metas extrínsecas superó el supuesto de homocedasticidad, como lo indicó la prueba de Levene, por lo que pueden ser utilizadas pruebas post hoc estándar con corrección de Tukey F(2,591)=1,56 p=0,169. Respecto del nivel educativo, como efecto principal, los resultados de estas pruebas revelaron que no hay diferencias significativas entre secundaria y pregrado (p=,242), pero sí se dieron entre secundaria y posgrado, así como entre pregrado y posgrado (p<,001 y p=,008, respectivamente). En cuanto a la interacción, los resultados muestran diferencias significativas en el sentido de que las mujeres del nivel de posgrado tienen medias mucho más bajas que las de las mujeres y los hombres en secundaria (p<,001 y p=,006, respectivamente) y que las presentadas por las mujeres en pregrado (p=,001).

Evidentemente, añadir una gráfica de líneas o barras que muestre los resultados contribuirá en gran medida a su lectura y comprensión. Obsérvese, por ejemplo, cómo pueden ser expresados los resultados referidos a las últimas dos escalas, si se añaden las gráficas correspondientes, como se evidencia en la figura 73.

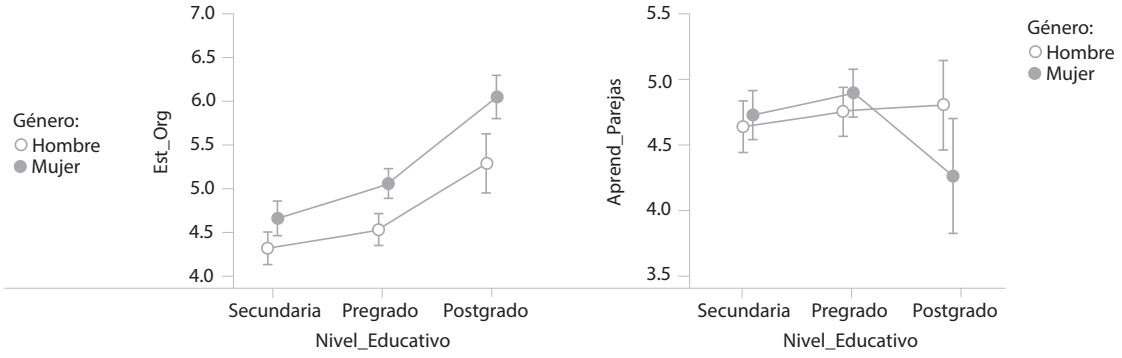


Figura 73. Organización y aprendizaje en parejas por sexo y nivel educativo

Respecto de la escala de organización, los resultados se presentan en la gráfica. Como se observa, las medias de esta escala muestran un comportamiento estrictamente creciente entre los niveles educativos en el que los puntajes de las mujeres muestran ser mayores que los de los hombres. Los efectos principales manifiestan ser significativos para el género, con tamaños de efecto entre pequeños y medianos F(1,591)=28,62 p<,001 η_p^2 =,05, así como para el nivel educativo, con tamaño de efecto entre medianos y grandes F(2,591)=38,17 p<,001 η_p^2 =,11. No hay efectos significativos asociados con la interacción entre género y nivel educativo F(2,591)=1,31 p=,271 η_p^2 <,01.

La prueba de Levene indicó la pertinencia de examinar pruebas post hoc de Games-Howell para la escala de organización F(5,591)=2,98 p=,011. Los resultados de las pruebas post hoc

revelaron diferencias significativas entre los niveles de secundaria y pregrado (p=,009), así como entre secundaria y posgrado y pregrado y posgrado (p<,001 en los dos casos).

Los resultados de la escala de aprendizaje en parejas aparecen en la gráfica. El Anova factorial mostró diferencias significativas ligadas a la interacción entre género y nivel educativo con tamaños de efecto pequeños F(2,591)=3,15 p=,044 =,01, pero no evidenció efectos principales significativos con género F(1,591)=1,00 p=,319 <,01, ni con nivel educativo F(2,591)=2,48 p=,085 <,01.

La prueba de Levene indicó la pertinencia de examinar pruebas post hoc estándar con corrección de Tukey F(5,591)=1,81 p=,109. Los resultados demostraron que la única diferencia significativa en la interacción entre género y niveles educativos se presenta entre las mujeres en pregrado, que muestran medias significativamente mayores a las mujeres en posgrado (p=,024).

Anova mixto

Presentación

Diseños mixtos (intra- e intersujetos)

Los *diseños mixtos de investigación* que involucran mediciones intrasujetos y mediciones intersujetos. Por su naturaleza, combinan las características de las medidas repetidas (con *i* repeticiones, con $i \geq 2$) y de las mediciones de contraste (entre *j* grupos, con $j \geq 2$).

El ejemplo más sencillo posible de este tipo de diseños es el diseño cuasiexperimental pretest/postest con grupo de control, que ya examinamos en el capítulo 8. Un diagrama de este diseño, siguiendo la notación de Campbell y Stanley (1961), es como sigue:

O	X	O
O		O

En este muy conocido y popular diseño existen dos medidas, pretest y postest, y, por tanto, tenemos la medición de la variable dependiente en dos momentos y, al tiempo, se han diferenciado dos grupos, experimental y de control, por la aplicación de una variable de tratamiento, o experimental (X).

En el capítulo 8 se examinaron las diferencias entre pretest y postest, mediante una prueba *t* de medidas apareadas y, por otro lado, se examinaron las diferencias entre los grupos experimental y control, mediante una prueba *t* de grupos independientes. Proceder de esta forma nos obligó a examinar cuatro pruebas diferentes y a interpretar los resultados de forma conjunta. Ahora podremos hacer el examen conjunto de todo el diseño cuasiexperimental con una sola prueba de hipótesis. Esta es la posibilidad que se abre con el análisis mixto de varianza.

Anova mixto

El *análisis mixto de varianza,* o *Anova mixto,* es una prueba de hipótesis paramétrica que resulta de la combinación entre el Anova de medidas repetidas (Anova MR) y el Anova factorial, que examinamos en el apartado anterior.

Como en el Anova factorial, en el Anova mixto se trabaja sobre todas las diferentes combinaciones de los factores, si bien ahora uno de los factores es intrasujetos, mientras que el otro es intersujetos. Esto hace que en el análisis del diseño cuasiexperimental pretest/postest deban ser consideradas cuatro combinaciones posibles, resultantes de la multiplicación entre los dos valores intrasujetos (pretest/postest) y los dos valores intersujetos (experimental/control) pretest del grupo experimental, pretest del grupo control, postest del grupo experimental y postest del grupo control.

Llamamos *efecto principal* al efecto de una de las variables independientes sobre la variable dependiente, ignorando los efectos de cualquier otra variable independiente. En el caso del diseño cuasiexperimental pretest/postest, tenemos dos efectos principales: la comparación de los datos entre pretest y postest, y la comparación de los datos entre los grupos experimental y de control. Se presenta una interacción entre los dos factores cuando un factor influye sobre el otro.

Como en el caso del Anova factorial, pueden distinguirse tres tipos de hipótesis nulas, con sus correspondientes hipótesis alternativas. La primera sobre los efectos intrasujetos, la segunda sobre los efectos intersujetos y la tercera sobre los efectos de las interacciones entre los dos factores.

En la medida en el que el Anova mixto es una prueba paramétrica que combina dos pruebas paramétricas, requiere también del cumplimiento de los supuestos combinados de las dos pruebas. Específicamente, el Anova mixto necesita del cumplimiento de los siguientes supuestos:

- El factor intersujetos deberá contener, al menos, dos niveles apareados.
- El factor intersujetos deberá contener, al menos, dos grupos independientes.
- La variable dependiente deberá ser métrica, continua y mostrar una distribución aproximadamente normal para todas las combinaciones posibles de los factores. Esto significa que debe ser aproximadamente simétrica y no mostrar valores atípicos muy significativos.
- Debe haber homogeneidad de varianza (homocedasticidad) para cada uno de los grupos. Si hay más de dos niveles, deberá haber *esfericidad* entre los grupos relacionados (véase Anova MR).

El supuesto de esfericidad solo se requiere para el caso en que el factor intrasujetos tenga tres, o más, niveles. Si se viola este supuesto, es posible utilizar correcciones al procedimiento. Las correcciones disponibles son la Greenhouse-Geisser y la Huynh-Feldt. Normalmente, se recomienda el uso de la corrección de Huynh-Feldt, pues ha mostrado ser la más eficiente y robusta entre las dos (Abdi, 2010).

En general, el Anova mixto produce dos tablas separadas: una referida a los efectos intrasujetos y otra, a los efectos intersujetos. Es importante interpretarlas teniendo a mano, al menos, un gráfico de las medias de la variable dependiente con sus distintas mediciones a lo largo del eje X, y diferenciando con líneas independientes el factor intersujetos.

En la tabla de los efectos intrasujetos se deben exponer en líneas sucesivas el efecto principal del factor intrasujetos y el efecto de su interacción con el factor intersujetos; en la línea final de la tabla se presenta la información sobre el residuo. En cada caso se enuncian, como en los otros Anova que se han examinado, la suma de cuadrados, los grados de libertad, el cuadrado medio, el valor del estadístico de prueba (F) y sus niveles de significación asociados (p). Si se han solicitado medidas de tamaño del efecto, esta tabla las incluirá para cada uno de estos dos efectos.

Por su parte, la tabla de efectos intersujetos presentará los mismos datos (suma de cuadrados, grados de libertad, cuadrado medio, *F*, *p* y las medidas de tamaño del efecto) de todos los factores intersujetos; la línea final presentará los datos de los residuos.

Con relación a las medidas de tamaño del efecto, debemos repetir las explicaciones y recomendaciones dadas en la sección el Anova MR. El programa JASP ofrece cuatro posibilidades para las medidas de tamaño del efecto: el *eta cuadrado* (η^2), el *eta parcial al cuadrado* (η_p^2), el *eta al cuadrado general* (η_G^2) y el *omega cuadrado* (ω^2). En general, se recomienda el uso de eta al cuadrado parcial (η_p^2), especialmente en muestras de buen tamaño (n>30). Si las muestras son pequeñas (n<30), se prefiere el uso del ω^2. El eta al cuadrado general (η_G^2), por su parte, es particularmente útil cuando existen algunos factores controlados activamente por el investigador y otros individuales no controlados.

La tabla 121 permite interpretar estas medidas de tamaño del efecto en el Anova mixto.

Tabla 121. Límites para la interpretación de las medidas de tamaño del efecto en el Anova mixto

Medida de tamaño del efecto	Nulo	Pequeño	Mediano	Grande
η^2	<0,1	0,1	0,25	0,37
η_p^2 (n>30)	<0,01	0,01	0,06	0,14
ω^2 (n<30)	<0,01	0,01	0,06	0,14

Fuente: Análisis estadístico con JASP.

Como en los Anova que ya hemos estudiado, el Anova mixto es una prueba global que señala diferencias en las medias, tanto en los efectos principales como en las interacciones, pero no indica entre cuales de los grupos específicos se presentan diferencias. Para saberlo, deben ser examinadas las pruebas *post hoc* apropiadas. Debe recalcarse, de nuevo, que las pruebas *post hoc* solo pueden ser examinadas en el caso en que el Anova mixto haya mostrado diferencias globales significativas.

En general, se recomienda el uso de las pruebas *post hoc* de Bonferroni. Esta opción, que resulta ser la más popular, es la única que está presente para este tipo de análisis en los dos paquetes que utilizamos.

Para el reporte de los resultados en un Anova mixto, se acostumbra el uso del siguiente formato, en cada uno de los efectos principales o las interacciones:

$$F(<gl1>, <gl2>)=<\text{Valor F}> p</= <\text{Valor p}> \eta^2_p/\omega^2=<\text{valor de tamaño del efecto}>$$

Cómo ejecutar un Anova mixto

Para correr el Anova mixto en el programa JASP, puede procederse a través del menú Anova/ Repeated Measures (recuadro 49). En el programa IBM-SPSS debe buscarse el procedimiento a través del "Modelo lineal general" (recuadro 50).

Recuadro 49. Cómo ejecutar un Anova mixto en JASP

/ANOVA/Repeated Measures ANOVA…
Debe digitarse el nombre global de factor y los nombres de cada una de las mediciones en el cuadro "Repeated Measures Factor". Esto permitirá trasladar las variables de las mediciones al cuadro "Repeated Measures Cells".
Debe pasarse el factor (o factores) intersujetos a la lista "Between Subject Factors".
Es recomendable seleccionar las siguientes opciones:
Display
√ Descriptive statistics
√ Estimates effect size
√ Partial η^2 √ ω^2 √ general η^2 (dependiendo del tamaño de la muestra)
Assumption Checks
√ Sphericity test (si procede, al haber más de tres niveles)
(es posible aquí seleccionar correcciones, dependiendo del test anterior)
Sphericity corrections
√ None o √ Huynh-Feldt
√ Homogeneity test
Post Hoc test
En este punto se pasa el factor, o la interacción, a la lista
√ Bonferroni
Display
√ Flag Significant Comparisions
Descriptive plots
Se pasa el factor de medidas repetidas a "Horizontal Axis" y el factor intersujeto a "Separate Lines". Si hay más factores, pueden pasarse a "Separate Plots".

Recuadro 50. Cómo ejecutar un Anova mixto en IBM-SPSS

/Analizar/Modelo lineal general/Medidas repetidas…
En este punto debe digitarse el nombre global de factor (por defecto será "factor 1") y se debe definir el número de niveles (n) y pulsar el botón "Añadir" y el botón "Definir". Esto dará paso a otro menú "Medidas repetidas". Allí deben ser tomadas las variables de los n niveles y pasadas a "Variables intrasujetos".

- En el botón "Gráficos" puede pasar de la lista "Factores" a "Eje horizontal" y "Añadir"

Pulsar "Continuar"

- En el botón "Post Hoc"… se selecciona la prueba adecuada

√ Tukey o √ Games-Howell

Pulsar "Continuar"

- En el botón "Opciones" es recomendable seleccionar:

√ Comparar los efectos principales. Allí arrastrar el nombre del factor a "Mostrar media para:" y seleccionar en la lista desplegable "Bonferroni"
√ Estadísticos descriptivos
√ Estimaciones del tamaño del efecto
Pulsar "Continuar"
Pulsar "Aceptar"

Ejemplo 1: evaluación del efecto de un programa de Matemáticas

En este ejemplo, examinaremos exactamente los mismos datos que trabajamos en el capítulo 10 (pruebas para dos medidas), en cuatro pruebas diferentes, pero ahora usando, como única prueba de hipótesis, el Anova mixto.

De acuerdo con el ejemplo, un profesor ha diseñado un programa pedagógico del área de Matemáticas y desea probar su efectividad. Para hacerlo, lo ha aplicado a uno de sus dos grupos escolares, mientras que el otro ha continuado con sus actividades usuales.

Este diseño de investigación se conoce como *diseño cuasiexperimental pretest/postest con grupo de control* (Campbell y Stanley, 1961), en el cual se ha definido, como variable experimental, la exposición de los estudiantes al nuevo programa de Matemáticas. Este diseño define un grupo "cuasiexperimental", en el que aplicará el programa, y uno "cuasicontrol" que continúa sus actividades usuales. Se dice que este diseño es "cuasiexperimental" y no "experimental", pues trabaja sobre grupos previamente conformados y, por lo tanto, no es posible asegurar su equiparabilidad (para hacerlo, los participantes deberían ser aleatoriamente asignados a cada grupo).

A fin de examinar las diferencias entre los grupos, se han aplicado pruebas de logro en la resolución de problemas matemáticos. Estas pruebas de logro producen información métrica a nivel de intervalo.

Planteamiento de las hipótesis

El Anova mixto admite la formulación de tres tipos de hipótesis nulas, y de sus correspondientes alternativas. La primera se refiere al efecto principal intrasujetos.

> *Hipótesis nula 1 (H_{0_1}). No hay diferencias significativas entre las medias del pretest y del postest en la prueba de Matemáticas.*
>
> *Hipótesis alternativa 1 (H_{a_1}). Existen diferencias significativas entre las medias de pre y postest de Matemáticas.*

La segunda se refiere al efecto principal intersujetos. Esto es, en este caso, a las diferencias entre los grupos cuasiexperimental y cuasicontrol.

> *Hipótesis nula 2 (H_{0_2}). No hay diferencias significativas en las medias de los grupos experimental y de control en la prueba de Matemáticas.*
>
> *Hipótesis alternativa 2 (H_{a_2}). Existen diferencias significativas entre los grupos experimental y de control en la prueba de Matemáticas.*

Finalmente, la tercera se refiere a los efectos de la interacción entre los dos factores.

> *Hipótesis nula 3 (H_{a_3}). No hay diferencias significativas entre los resultados de la prueba de Matemáticas de los grupos cuasiexperimental y cuasicontrol a través del pretest y del postest.*
>
> *Hipótesis alternativa 3 (H_{a_3}). Existen diferencias significativas entre los resultados de la prueba de Matemáticas de los grupos cuasiexperimental y cuasicontrol a través del pretest y del postest.*

Selección de la prueba

El diseño de investigación plantea un factor intrasujetos de dos niveles y un factor intersujetos de dos niveles. La opción ideal, para este tipo de diseño es la prueba de hipótesis del análisis mixto de varianza (Anova mixto). Debe anotarse que no existe un equivalente no paramétrico de esta prueba.

La prueba requiere de una variable dependiente numérica y continua, un factor intrasujetos y un factor intersujetos. Estas tres condiciones se cumplen. La variable dependiente son los resultados de una prueba de Matemáticas. Existe un factor intrasujetos marcado por dos aplicaciones de la prueba, y un factor intersujetos que señala dos grupos (experimental y control).

Además de lo anterior, el Anova mixto requiere de la distribución aproximadamente normal de la variable dependiente entre los grupos y la homocedasticidad de la variable para cada grupo. Respecto de la normalidad, este supuesto se cumple, de acuerdo con los resultados de las pruebas de Shapiro-Wilk de la tabla 48 (capítulo 10). De igual forma, el presupuesto de homocedasticidad se cumple, tanto para el pretest como para el postest, de acuerdo con los resultados de la prueba de Levene de la tabla 49 (capítulo 10). En síntesis, se cumplen todos los supuestos de la prueba.

Cálculo de resultados

Los resultados generales de las medias por cada una de las pruebas de forma diferenciada para cada grupo se presentan en la figura 74 y en la tabla 122. La gráfica muestra que, para el momento del pretest, los grupos experimental y de control expresan medias en la prueba de Matemáticas muy similares. Ya para el momento del postest, la media de la prueba de Matemáticas en el grupo experimental parece ser bastante mayor que la media de la prueba en el grupo de control. Las barras alrededor de las medias indican que los intervalos de confianza de las dos medias se superponen levemente. Los resultados de las pruebas nos indicarán la significación de la diferencia.

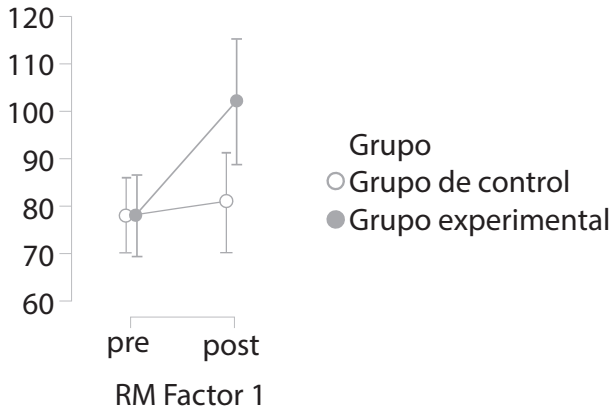


Figura 74. Puntaje en la prueba de Matemáticas por prueba y grupo

Tabla 122. Estadísticos del puntaje en la prueba de Matemáticas, por prueba y grupo

RM factor 1	Grupo	Media	DE	N
Post	Grupo de control	81,31	26,12	25
	Grupo experimental	102,32	31,94	25
Pre	Grupo de control	78,37	18,97	25
	Grupo experimental	78,27	20,88	25

El Anova mixto produce tres diferentes tablas. La primera indica los efectos intrasujetos. Como lo enseña la tabla, existe un efecto principal significativo (p<,001), que muestra diferencias entre el pretest y el postest, con un tamaño del efecto grande. Por otro lado, la prueba muestra también un efecto significativo de la interacción entre los dos factores (p=,001) con un tamaño de efecto grande. Esto podría indicar diferencias entre los dos grupos para el nivel del postest, pero no para el nivel del pretest (tabla 123).

Tabla 123. Tabla del Anova mixto. Efectos intrasujeto

Casos	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
RM factor 1 * grupo	2783.046	1	2783.046	11.488	0.001	0.193
Residuos	11628.025	48	242.251			

Nota: suma de los cuadrados tipo III.

La tabla 124 muestra el efecto principal de la diferencia ligada al factor intersujetos. Como se observa, no se verifica un efecto principal significativo que muestre diferencias entre el grupo experimental y el de control (p=,106). El tamaño de efecto se encuentra entre pequeño y mediano.

Tabla 124. Tabla de del Anova mixto. Efectos intersujeto

Casos	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
Residuos	48354.459	48	1007.385			

Nota: suma de los cuadrados tipo III.

Dado que la prueba mostró efectos generales significativos de la interacción de los dos factores, podemos examinar las pruebas *post hoc* de la interacción. Los resultados de estas pruebas, con la corrección de Bonferroni, se presentan en la tabla 125.

Tabla 125. Pruebas post hoc para la verificación de diferencias en la interacción grupo-factor (prueba pretest-postest)

Comparaciones *post hoc*–grupo * RM factor 1					
		DE	SE	t	p_{bonf}
Grupo de control, pre.	Grupo, experimental, pre.	0.097	7.070	0.014	1.000
	Grupo de control, post.	-2.944	4.402	-0.669	1.000
	Grupo experimental, post.	-23.950	7.070	-3.387	0.007 **
Grupo experimental, pre.	Grupo de control, post.	-3.041	7.070	-0.430	1.000
	Grupo experimental, post.	-24.046	4.402	-5.462	< .001 ***
Grupo de control, post.	Grupo experimental, post.	-21.005	7.070	-2.971	0.024 *

Nota: el valor de *p* y los intervalos de confianza fueron ajustados para comparar una familia de 6 estimados (los intervalos de confianza fueron corregidos con el método Bonferroni).

* p < .05, ** p < .01, *** p < .001

Como se observa, las pruebas muestran diferencias significativas entre el pretest y el postest del grupo experimental (p<,001) así como entre el postest del grupo experimental y el del grupo de control (p=,024). Este último punto es crucial para la verificación del efecto del programa.

Interpretación y reporte

Los resultados podrían ser reportados como sigue:

> *Los resultados del Anova mixto indicaron que la media de la prueba de Matemáticas en el postest muestra ser significativamente mayor que la media del pretest con un tamaño del efecto grande F(1,48)=18,80 p<,001 η^2_p=0,28.*
>
> *Aunque no se verifican efectos principales significativos ligados a las diferencias entre los grupos experimental y control F(1,48)=2,71 p=,106 η^2_p=0,05, se presenta una interacción significativa, con un tamaño de efecto grande, entre el grupo y el momento de la evaluación F(1,48)=11,49 p=,001 η^2_p=0,19. Al respecto, el análisis post hoc con la corrección de Bonferroni mostró que, aunque no hay diferencias significativas en las medias de los dos grupos en el pretest (p=1,000), las medias del postest en el grupo experimental son significativamente mayores que las del pretest en el mismo grupo (p<,001), y significativamente mayores que las medias del postest en el grupo de control (p=,024).*
>
> *En conclusión, los puntajes de Matemáticas en el grupo experimental se incrementaron de forma significativa respecto de sus condiciones iniciales y respecto de sus compañeros del grupo control. Esto debe ser atribuido a un efecto del programa experimental.*

Ejemplo 2: Seis meses de aprendizaje cooperativo sobre la lectura y la escritura

Este segundo ejemplo, basado en un caso real de investigación, hace más extremas las dificultades de un diseño cuasiexperimental.

Una profesora desea examinar el efecto del aprendizaje cooperativo para el logro en la asignatura de Español del cuarto grado. Para hacerlo, a uno de sus dos grupos de cuarto grado lo lleva en una situación pedagógica de aprendizaje cooperativo en la asignatura durante seis meses completos, mientras que en el otro se continúa con las actividades usuales. En los dos grupos se aplican dos pruebas de comprensión de lectura y producción escrita (prueba de Español). Dos versiones de esta prueba se aplican, en los dos grupos, al inicio (pretest) y al final (postest). Interesa examinar el efecto del aprendizaje cooperativo sobre la comprensión lectora y la producción escrita durante ese periodo.

Formulación de las hipótesis

Como sabemos, el Anova mixto admite la formulación de tres tipos de hipótesis nulas, y de sus correspondientes alternativas. La primera se refiere al efecto principal intrasujetos.

> *Hipótesis nula 1 (H_{0_1}). No hay diferencias significativas entre las medias del pretest y del postest en la prueba de Español.*
>
> *Hipótesis alternativa 1 (H_{a_1}). Existen diferencias significativas entre las medias de pretest y postest de español.*

La segunda hipótesis se refiere al efecto principal intersujetos. Esto es, en este caso, a las diferencias entre los grupos cuasiexperimental y cuasicontrol:

> *Hipótesis nula 2 (H_{0_2}). No hay diferencias significativas en las medias de los grupos experimental y de control en la prueba de Español.*
>
> *Hipótesis alternativa 2 (H_{a_2}). Existen diferencias significativas entre los grupos experimental y de control en la prueba de Español.*

Finalmente, la tercera se refiere a los efectos de la interacción entre los dos factores:

> *Hipótesis nula 3 (H_{0_3}). No hay diferencias significativas entre los resultados de la prueba de Español de los grupos cuasiexperimental y cuasicontrol a través del pretest y del postest.*
>
> *Hipótesis alternativa 3 (H_{a_3}). Existen diferencias significativas entre los resultados de la prueba de Español de los grupos cuasiexperimental y cuasicontrol a través del pretest y del postest.*

Puede ser importante notar que, para nuestro caso, la hipótesis que más interés tiene es la tercera, relacionada con la interacción de los dos factores. Con la primera, la hipótesis intrasujetos, solo podemos verificar diferencias globales entre los dos momentos, pero no podemos verificar diferencias entre los grupos atribuibles al tratamiento. Con la segunda, la hipótesis intersujetos, solo podremos verificar diferencias globales entre los dos grupos, diferencias que podrían ser iguales en el pretest y en el postest. Solo con la tercera hipótesis podemos verificar un efecto significativo del tratamiento que se diferencie entre los dos grupos.

Selección de la prueba

Como opción principal, en este caso, debemos seleccionar la prueba de hipótesis del análisis mixto de varianza.

La prueba requiere de una variable dependiente numérica y continua, un factor intrasujetos y un factor intersujetos. Estas tres condiciones se cumplen: 1) la variable dependiente son los resultados de la prueba de Español; 2) existe un factor intrasujetos marcado por dos aplicaciones de la prueba (pretest y postest), y 3) existe un factor intersujetos que señala dos grupos (experimental y control).

Además de lo anterior, el Anova mixto requiere que la variable dependiente tenga una distribución aproximadamente normal entre los grupos y que presente homocedasticidad para cada grupo. Al respecto de la normalidad aproximada, este supuesto se cumple.

Con relación al supuesto de homocedasticidad, los resultados de las pruebas de Levene indican que debe rechazarse la hipótesis nula que señala la igualdad de varianzas, para las dos medidas de la variable dependiente (tabla 126).

Tabla 126. Pruebas de Levene de igualdad de varianzas para el pretest y el postest de Español

	F	gl1	gl2	p
Pretest de Español	8,696	1	69	0,004
Postest de español	4,556	1	69	0,036

Esta situación debe alertarnos acerca de la posibilidad de incrementar el error de tipo I en este caso. Lamentablemente, no tenemos correcciones ni alternativas a la situación, por lo que debemos proseguir, haciendo la advertencia.

Cálculo de resultados

La figura 75 muestra las medias del pretest y del postest de forma diferenciada para cada uno de los grupos. Como se observa, existe una diferencia notoria en los valores de las medias en el pretest entre los grupos, mientras que en el postest estas medias parecen haberse vuelto más similares (tabla 127). Esto permitiría suponer tres puntos de importancia para el experimento.

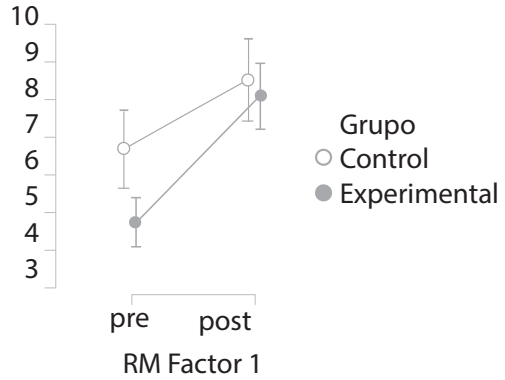


Figura 75. Medias de la prueba de Español en pretest y postest por grupo

Tabla 127. Estadísticos descriptivos del pretest y postest de Español en el grupo experimental y de control

Descriptivo				
RM Factor 1	Grupo	M	DE	n
Post	Control	8.487	3.433	39
	Experimental	8.063	2.501	32
Pre	Control	6.615	3.241	39
	Experimental	4.625	1.913	32

El primer punto, es que parece claro que los dos grupos no eran completamente equiparables respecto de sus condiciones de competencia inicial en el área de Español. Este es uno de los riesgos de seleccionar un diseño cuasiexperimental que, en este caso, parece confirmarse.

El segundo punto de importancia es que, para los dos grupos, parece mostrarse un incremento claro de los promedios de las pruebas. Este efecto es normal, relativamente esperable, y puede ser explicado como un resultado natural de procesos de maduración y del aprendizaje.

El tercer punto es el hecho de que la situación de salida, marcada por el postest, parezca indicar una relativa similaridad entre los dos grupos. Dadas las diferencias iniciales entre los dos grupos, esta relativa equiparación podría señalar que el tratamiento al que fue sometido el grupo cuasiexperimental parece mostrar un mayor efecto, sobre los resultados del postest, que el correspondiente del grupo cuasicontrol.

Los resultados del Anova mixto se presentan en las tablas 128 a 130. Iniciando con la tabla 128, de las diferencias intrasujetos, los resultados indican diferencias principales significativas entre las dos pruebas (p<,001), que revelan una mayor media en el postest frente al pretest. Esto era esperable como un efecto natural de la maduración, el aprendizaje y la historia ocurrida entre los dos momentos.

Tabla 128. Anova mixto. Diferencias intrasujetos

Casos	Suma de cuadrados	gl	Cuadrado medio	F	p	η^2_p
RM Factor 1	247 742	1	247 742	53 400	< ,001	0,436
RM Factor 1 * GRUPO	21 545	1	21 545	4644	0,035	0,063
Residuos	320 117	69	4639			

Nota: suma de los cuadrados tipo III.

Por otro lado, la tabla 129, de los efectos intrasujetos, también señala una diferencia significativa ligada a la interacción del factor intrasujetos con el factor intersujetos (p=,035). Para verificar los grupos específicos que muestran diferencias deberán ser examinadas las pruebas *post hoc.* La tabla muestra que existen diferencias principales significativas entre los dos grupos (p=,043). El examen de estas diferencias indica que el grupo cuasicontrol muestra mayores medias en la prueba de Español que el grupo cuasiexperimental, en gracia a sus superiores resultados en el pretest.

Tabla 129. Anova mixto. Diferencias íntersujetos

Casos	Suma de cuadrados	gl	Media cuadrática	F	p	η^2_p
grupo	51 261	1	51 261	4240	0,043	0,058
Residuos	834 232	69	12 090			

Nota: suma de los cuadrados tipo III.

Dado que la interacción entre el factor intrasujetos y el factor intersujetos mostró diferencias globales significativas, pueden examinarse las pruebas *post hoc* que verifiquen las diferencias específicas. La tabla 129 muestra los resultados de estas pruebas.

Tabla 130. Anova mixto. Pruebas post hoc de las interacciones entre el grupo y la prueba (pretest-postest)

Comparaciones *post hoc*–GRUPO * RM Factor 1							
			IC 95% para DE				
		Diferencia media	Más bajo	Más alto	SE	t	p_{bonf}
Control, pre	Experimental, pre	1,990	0,138	3,842	0,690	2,885	0,028 *
	Control, post	-1,872	-3,197	-0,547	0,488	-3,837	0,002 **
	Experimental, post	-1,447	-3,299	0,405	0,690	-2,098	0,229
Experimental, pre	Control, post	-3,862	-5,714	-2,010	0,690	-5,599	< ,001 ***
	Experimental, post	-3,438	-4,900	-1975	0,538	-6,384	< ,001 ***
Control, post	Experimental, post	0,425	-1,427	2,277	0,690	0,616	1,000

Nota: el valor de *p* y los intervalos de confianza fueron ajustados para comparar una familia de 6 estimados (los intervalos de confianza fueron corregidos con el método Bonferroni).

* p < .05, ** p < .01, *** p < .001.

Los resultados contenidos en la tabla 129 muestran que, primero, existen diferencias significativas entre el pretest y el postest, tanto en el grupo experimental (p<,001) como en el grupo de control (p=,002); segundo, que, a pesar de que existan diferencias significativas entre los dos grupos en el pretest, a favor del grupo de control (p=,028), no hay diferencias significativas entre los grupos en el postest (p=1,000). En este sentido, se puede concluir que la exposición al aprendizaje cooperativo contribuyó a eliminar la diferencia inicialmente presente entre los dos grupos, equiparándolos, por lo que se puede concluir un efecto significativo del aprendizaje cooperativo sobre los resultados de la prueba de Español.

Interpretación y reporte

Utilizando las convenciones ya expuestas para la presentación de los resultados en texto, el reporte puede quedar de la siguiente forma:

> *Se examinó un análisis mixto de varianza en que se usaron como variable dependiente los resultados de la prueba de Español tomados antes y después de la aplicación del programa de aprendizaje cooperativo. Los resultados indicaron 1) diferencias significativas entre el pretest y el postest con un tamaño del efecto grande $F(1,69)=53,40$ $p<,001$ $\eta^2_p=0,44$; 2) diferencias principales entre los dos grupos, a favor del grupo de control $F(1,69)=4,24$ $p=,043$ $\eta^2_p=0,6$; y 3) diferencias significativas ligadas a la interacción de los dos factores $F(1,69)=4,64$ $p=,035$ $\eta^2_p=0,06$.*
>
> *Las pruebas post hoc con corrección de Bonferroni indicaron diferencias significativas entre el pretest y el postest, tanto en el grupo experimental ($p<,001$) como en el grupo de control ($p=,002$). De igual forma, revelaron que, a pesar de que existían diferencias significativas entre los dos grupos en el pretest, a favor del grupo de control ($p=,028$), no hay diferencias significativas entre los dos grupos en el postest ($p=1,000$). En este sentido, se puede concluir que la exposición al aprendizaje cooperativo contribuyó a eliminar la diferencia inicialmente presente entre los dos grupos, equiparándolos. Esto demuestra que se puede concluir un efecto significativo del aprendizaje cooperativo sobre los resultados de la prueba de Español.*

Análisis de covarianza (Ancova)

Presentación

El *análisis de la covarianza*, o Ancova, es una combinación del Anova y de la regresión lineal múltiple. Su nombre es un acrónimo del inglés (*analisys of covariance*). Fue creado por R. Fisher y fue publicado por primera vez en 1947 en *Biometrics*, en un trabajo titulado "The analysis of covariance method for the relation between a part and the whole". Es un procedimiento estadístico que permite eliminar la heterogeneidad causada en la variable dependiente por la influencia de una o más variables cuantitativas, conocidas como *covariables*, que están relacionadas con la variable dependiente, pero cuyo efecto se pretende "separar" del análisis.

En general, en la investigación educativa y social resulta muy importante controlar las fuentes de error experimental, que aparecen debido a variables extrañas, ya que nos pueden alterar los resultados y dificultar la interpretación de los efectos que estamos buscando. Las técnicas que permiten hacerlo pueden ser de dos tipos: *a priori*, que se utilizan antes de aplicar los tratamientos y de recoger los datos, y *a posteriori*, que se utilizan una vez obtenidos los datos.

Siempre es preferible el uso de técnicas *a priori* para la reducción de error experimental. Sin embargo, con el uso de este tipo de técnicas no siempre es posible controlar todas las fuentes de error; en estas ocasiones debemos utilizar técnicas *a posteriori*, tales como el análisis de covarianza.

En esencia, al suprimir el efecto de la covariable, el análisis de covarianza consigue eliminar la heterogeneidad de la variable que estamos estudiando. Esto hace que, cuantas más covariables tengamos, menos variabilidad tendrán los datos y por tanto más potencia estadística tendrá la prueba. Como se recordará, la potencia estadística es la probabilidad de que una prueba identifique correctamente el impacto que tiene un tratamiento en los resultados que estamos estudiando.

El resultado que nos da el análisis de covarianza es una puntuación corregida a la que se le ha restado la cuantía o el valor atribuible a las covariables. Así, el análisis de covarianza permite aumentar la precisión de los experimentos al eliminar los efectos de variables que no tienen nada que ver con los tratamientos pero que, sin embargo, sí están influyendo en los resultados.

Salvo por lo relacionado con las covariables, los supuestos del Ancova son idénticos a los del Anova factorial. Esto es, se requiere:

- Observaciones aleatoriamente seleccionadas y, por tanto, independientes entre sí.
- Una, o varias, variables categoriales que actuarán como variables independientes (factores).
- Una variable dependiente métrica y continua.
- La variable dependiente debe mostrar una distribución aproximadamente normal en todas poblaciones definidas por el cruce de los factores.
- La variable dependiente debe mostrar homocedasticidad entre las poblaciones definidas por el cruce de los factores.

En teoría, no hay límites para el número de covariables por considerar en un Ancova. En la práctica, sin embargo, rara vez se exceden dos o tres covariables. Para que el Ancova tenga sentido, las covariables deben cumplir con los siguientes supuestos:

- *Las covariables no deben estar relacionadas con los factores independientes*. Esto puede ser verificado mediante un Anova en que la covariable sea considerada como variable dependiente.
- *Cada una de las covariables debe presentar una relación lineal significativa con la variable dependiente*. Esto puede ser verificado mediante una regresión simple. Esta regresión es presentada en la prueba misma.
- *Las pendientes de las rectas de regresión en todos los grupos deben ser iguales*. Esto puede ser verificado al incluir el examen de la interacción entre la covariable y el factor independiente en el modelo del Ancova. Sin embargo, para el examen final del modelo conviene eliminar esta interacción que, en general, no interesa.

Los anteriores supuestos referidos a la covariable son centrales para este análisis. La ausencia de cumplimiento de cualquiera de ellos invalida el procedimiento por cuanto conduce a una situación en la que la covariable no mejora la capacidad explicativa del modelo.

Los resultados del Ancova son prácticamente iguales a los del análisis factorial de varianza en casi todo, excepto en el cálculo del efecto de la covariable. Este efecto será expresado como una línea más en la tabla de las diferencias intersujetos, pero será sustraído de la variable dependiente para el cálculo de los efectos de los factores independientes.

Para el cálculo del tamaño del efecto, están disponibles las medidas usuales del Anova: eta cuadrado (η^2), eta cuadrado parcial (η^2_p) y el omega al cuadrado (ω^2). Como en los casos anteriores se recomienda el uso del eta al cuadrado parcial, salvo en los casos en que la muestra es pequeña ($n<30$), en donde se recomienda el omega cuadrado.

Cómo ejecutar un Ancova

Para examinar un análisis de covarianza en los diferentes programas puede procederse de la forma que se presenta en el recuadro 51 para el programa JASP. En el programa IBM-SPSS debe buscarse el procedimiento a través del menú "Modelo lineal general/Univariado…" (recuadro 52).

Recuadro 51. Cómo ejecutar una Analisis de Covarianza (Ancova) en JASP

/ANOVA/ANCOVA

En este punto debe pasarse la variable dependiente a la lista "Dependent variable", los factores independientes a la lista "Fixed factors" y las covariables a la lista "Covariantes"

- Display
  √ Descriptive statistics
  √ Estimates effect size
  √ Partial η_p^2 o √ ω^2 (dependiendo del tamaño de la muestra)
- Model. En este punto estarán la variable independiente y la covariable en la lista "Model terms". Para examinar el supuesto de interacción entre estas dos, la interacción debe ser incluida señalando las dos variables y pasando esta interacción a la lista. Para el modelo final, esta interacción debe ser eliminada
- Assumption Checks
  √ Homogeneity test
  √ Q-Q plot of residuals
- Post Hoc test

(dependiendo de la homogeneidad, se selecciona la prueba adecuada)

√ Tukey o √ Games-Howell

- Descriptive Plots (pasar la variable de factor a "Horizontal Axis")
  √ Display error bars

Recuadro 52. Cómo ejecutar una Analisis de Covarianza (Ancova) en IBM-SPSS

/Analizar/Modelo lineal general/Univariante…

En este punto debe pasarse la variable dependiente a la casilla "Variable dependiente", los diferentes factores a la lista "Factores fijos" y las covariables a la lista "Covariables".

- En el botón "Modelo"

  Pueden especificarse los términos del modelo. Se sugiere incluir todos los efectos principales seleccionar la variable dependiente y la covariable y pasarlas juntas a la lista "Modelo" para verificar el supuesto de la ausencia de interacción. Verificado el supuesto puede eliminarse del modelo
- En el botón "Gráficos"

  Pasar de la lista "Factores" a un factor a "Eje horizontal" y otro a "Líneas separadas". Es posible pasar un tercer factor a "Gráficos separados"

  Pulsar "Continuar"
- En el botón "Post hoc"…

  Deben seleccionarse los factores sobre los que se desean la pruebas *post hoc* y seleccionar la prueba adecuada.

  √ Tukey o √ Games-Howell

  Pulsar "Continuar"
- En el botón "Opciones":
  √ Estadísticos descriptivos
  √ Pruebas de homogeneidad
  √ Estimaciones del tamaño del efecto

  Pulsar "Continuar"

Pulsar "Aceptar"

Ejemplo. Cómo controlar el efecto del estilo cognitivo

Este ejemplo está basado en el mismo caso de investigación que examinamos en el Anova mixto, e ilustra claramente el potencial del Ancova para determinar el efecto de un tratamiento.

Una profesora desea examinar el efecto de la metodología conocida como *aprendizaje cooperativo* para el logro en la asignatura de Matemáticas del quinto grado. Para hacerlo, divide a su grupo de estudiantes en dos partes, utilizando una estrategia aleatoria: a los estudiantes con número impar en la lista, los asigna al grupo experimental, mientras que a los estudiantes de número par en la lista los asigna al grupo de control. Como los grupos son equiparables, en gracia al proceso de asignación aleatoria, no se requiere usar un pretest. Esto se conoce como diseño experimental postest. Deben compararse simplemente las puntuaciones del postest.

Esta profesora ha observado previamente y conoce acerca de las relaciones entre el estilo cognitivo, en la dimensión de dependencia-independencia de campo (DIC) y el logro en Matemáticas y quisiera aislar y separar este efecto, del efecto propio de la estrategia del aprendizaje cooperativo. Para hacerlo, aplica a los dos grupos la prueba de estilo cognitivo (EFT). Esta prueba produce un puntaje, entre 0 y 50 puntos, que, posiblemente, estará positivamente relacionado con el postest de Matemáticas. Se pretende ahora utilizar un análisis de covarianza para examinar el efecto del aprendizaje cooperativo eliminando el posible efecto del estilo cognitivo.

Planteamiento de las hipótesis

Las hipótesis correspondientes al Ancova son idénticas a las del Anova factorial, después de excluir los efectos de la(s) covariable(s).

> *Hipótesis nula (H_0). No existen diferencias significativas en las medias de la escala de logro en Matemáticas entre los grupos de tratamiento, una vez ha sido excluido el efecto del estilo cognitivo.*
>
> *Hipótesis alternativa (H_a). Existen diferencias significativas en las medias de la escala de logro en Matemáticas entre los grupos de tratamiento, una vez ha sido excluido el efecto del estilo cognitivo.*

En este caso solo hemos teniendo en cuenta un factor independiente. De haber considerado más factores, deberíamos plantear hipótesis para cada uno de los efectos principales, y para las diferentes interacciones entre los mismos.

Selección de la prueba

La selección del Ancova está sujeta al cumplimiento de los supuestos propios del Anova, además del cumplimiento de tres supuestos específicos relacionados con las covariables.

Iniciando con los supuestos del Anova, se requiere de una variable dependiente métrica con una distribución aproximadamente normal. La figura 76 confirma este primer supuesto.

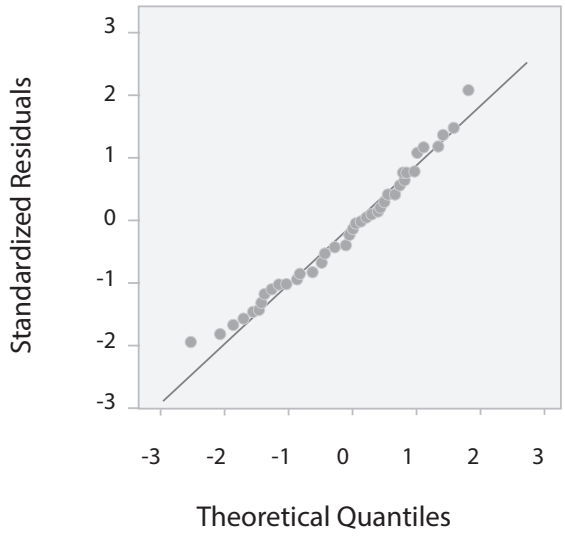


Figura 76. Gráfica Q-Q del postest de Matemáticas

El segundo de los supuestos del Anova es la homocedasticidad de la variable dependiente entre los grupos de tratamiento. La tabla 131 muestra los resultados de la prueba de Levene, que confirma la relativa igualdad de las varianzas entre los grupos.

Tabla 131. Prueba de Levene para el postest de Matemáticas por grupo

F	gl1	gl2	p
1,187	1,000	67,000	0,280

Los supuestos específicos que deben ser cumplidos por las covariables son tres. Primero, que las covariables no deben estar relacionadas con los factores independientes. Esto puede ser fácilmente verificado mediante un Anova en el que la covariable sea considerada como variable dependiente y el factor, la variable independiente. Existen dos supuestos más para la covariable. Al examinar esta prueba, los resultados indican que no puede rechazarse la hipótesis nula; esto es, no hay un efecto significativo del grupo sobre el puntaje de estilo cognitivo $F(1,72)=1,65$ $p=,203$ $\eta^2_p=0,02$. En esta medida, el primer supuesto se cumple.

El segundo supuesto es que cada una de las covariables debe presentar una relación lineal significativa con la variable dependiente y el tercer supuesto es que las pendientes de las rectas de regresión en todos los grupos deben ser iguales. Estos dos supuestos pueden ser examinados al correr una primera versión del Ancova en la cual se ha incluido el examen de la interacción entre la variable independiente y la covariable. Cuando esto se hace, para el caso del ejemplo, los resultados son como aparecen en la tabla 132.

Tabla 132. Tabla del Ancova incluyendo la interacción entre la covariable y la variable independiente

Ancova–matemáticasb						
Casos	**Suma de cuadrados**	**gl**	**Media cuadrática**	**F**	**p**	**η^2_p**
Grupo	21.519	1	21.519	3.978	0.050	0.058
eft	33.297	1	33.297	6.155	0.016	0.087
grupo * eft	6.614	1	6.614	1.223	0.273	0.018
Residuos	351.626	65	5.410			

Nota: suma de los cuadrados tipo III.

Respecto del segundo supuesto, de acuerdo con lo registrado en la tabla, el frente del nombre de la covariable (eft) se indica una relación significativa entre la covariable y la variable dependiente F(1,65)=6,16 p=,016 η^2_p=,09. Esto verifica el supuesto de la relación lineal significativa entre la covariable y la variable dependiente.

El tercero de los supuestos de las covariables es que las pendientes para todos los grupos deben ser iguales. Este supuesto se verifica constatando la ausencia de una interacción significativa entre la covariable y los factores independientes. Como se observa en la tabla 132, la interacción entre grupo y la covariable eft no muestra ser significativa *F*(1,65)=1,22 *p*=,273 η^2_p=,02, lo que permite verificar este tercer supuesto.

Los resultados, hasta el momento, permiten constatar la verificación de los supuestos del Ancova. La prueba misma, sin embargo, aún no ha sido examinada, debido a que los datos contenidos en la tabla incluían el examen del efecto de la interacción entre la variable independiente y la covariable, lo cual modifica el examen del efecto de la variable independiente sobre la dependiente. Examinaremos los resultados del Ancova en la siguiente sección.

Cálculo de resultados

La gráfica de la figura 77 muestra las diferencias en las medias de prueba de Matemáticas entre los dos grupos. Como se observa, la media de la prueba en el grupo experimental es bastante más alta que en el grupo de control.

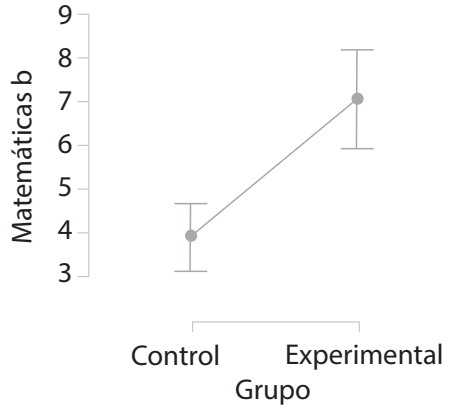


Figura 77. Puntaje en el postest de Matemáticas por grupo

Tabla 133. Descriptivos del postest de Matemáticas por grupo

Descriptivos–MATEMÁTICASb			
Grupo	**Media**	**DE**	**N**
Control	4.378	2.019	37
Experimental	7.188	2.776	32

La tabla muestra los resultados del Ancova. Primero, debe observarse que los resultados indican una relación lineal significativa entre la covariable (EFT) y la variable dependiente (p=,028). Esto verifica el último de los supuestos de la prueba.

Tabla 134. Resultados del Ancova del postest de Matemáticas por grupo, controlando el efecto del puntaje EFT

Ancova–MATEMÁTICASb						
Cases	**Suma de cuadrados**	**gl**	**Media cuadrática**	**F**	**p**	**η^2_p**
Grupo	152.244	1	152.244	28.049	< .001	0.298
EFT	27.338	1	27.338	5.037	0.028	0.071
Residuos	358.239	66	5.428			

Nota: suma de los cuadrados tipo III.

Por otro lado, el Ancova muestra diferencias significativas entre los dos grupos (p<,001), excluido el efecto del estilo cognitivo.

Interpretación y reporte

Para el reporte de los resultados del Ancova se sigue el formato que hemos venido utilizando para los diferentes Anova:

F(gl1,gl2)=<valor F> p=<valor p> η^2_p=<valor eta cuadrado parcial>

Siguiendo este formato, los resultados pueden quedar expresados de la siguiente forma:

> *Para la verificación de las diferencias en los niveles de aprendizaje matemático, entre los grupos de aprendizaje cooperativo y aprendizaje tradicional, manteniendo controlado el efecto del estilo cognitivo sobre el logro en Matemáticas, se examinó un análisis de covarianza (Ancova) sobre los resultados del postest de Matemáticas, definiendo como covariable el puntaje en la prueba de estilo cognitivo (EFT). Los supuestos del Ancova fueron adecuadamente satisfechos; en particular se observa 1) que no hay un efecto significativo del grupo sobre el puntaje de estilo cognitivo F(1,72)=1,65 p=,203 η^2_p=0,02; 2) que existe una relación lineal significativa entre la prueba de estilo cognitivo y el puntaje del postest de Matemáticas con un tamaño del efecto entre intermedio y grande F(1,66)=27,34 p=,028 η^2_p=0,07; y 3) que no hay una interacción significativa entre la covariable y el grupo F(1,65)=1,22 p=,273 η^2_p=,02.*
>
> *Los resultados del Ancova indicaron que, después de controlar el efecto del estilo cognitivo, se observa una diferencia muy significativa entre los dos grupos de tratamiento, a favor del grupo experimental, con un tamaño del efecto grande F(1,66)=152,24 p<,001 η^2_p=0,30.*

Referencias

Referencias a temas estadísticos

Abdi, H. (2010). The Greenhouse-Geisser Correction. En N. Salkind (Ed.), *Encyclopedia of Research Design*. Sage.

Anastasi, A. y Urbina, S. (1998). *Test psicológicos*. (7.ª ed.). Prentice Hall.

Aron, A. y Aron, E. (2002). *Estadística para psicología.* Prentice Hall.

Campbell, D. y Stanley, J. (1966). *Diseños experimentales y cuasiexperimentales en la investigación social*. Amorrortu.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2. Auflage).* Erlbaum.

Cohen, J. (1994). The Earth is round (p<0.05). *American Psychologist, 49*(12), 155-159.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Domínguez-Lara, S. A. y Merino-Soto, C. (2015). ¿Por qué es importante reportar los intervalos de confianza del coeficiente alfa de Cronbach? *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 13*(2), 1326-1328.

Ebel, R. L. (1965). Confidence weighting and test reliability. *Journal of Educational Measurement, 2*(1), 49-57. https://doi.org/10.1111/j.1745-3984.1965.tb00390.x

Fisher, R. A. (1947). The analysis of covariance method for the relation between a part and the whole. *Biometrics, 3*, 65-68.

Fritz, C., Morris, P. y Richler, J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology General, 141*(1), 2-18· http/doi.org/10.1037/a0024338

Frost, J. (2017). *Benefits of Welch's Anova compared to the classic one way Anova.* Statistics By Jim. https://statisticsbyjim.com/anova/welchs-anova-compared-to-classic-one-way-anova/

García, F. y Musitu, G. (2014). *Autoconcepto Forma 5.* TEA Ediciones.

Goss-Sampson, M. (2019). *Statistical analysis in JASP: A guide for students.* http/doi.org/10.6084/m9.figshare.9980744

Hair, J.; Anderson, R. (Comps.). (2004). *Análisis multivariante* (R. Tatham, y W. Black trads.). Prentice Hall. (Obra original publicada en 1999).

Hunter, J. E. (1997). Needed: A band on the significance test. *Psychological Science, 8*(1), 3-7. https://doi.org/10.1111/j.1467-9280.1997.tb00534.x

Jensen, A. R. (1980). *Bias in mental testing.* The Free Press.

Kim, H. Y. (2017). Statistical notes for clinical researchers and Fisher's exact test. *Restorative Dentistry & Endodontics, 42*(2), 152-155.

Kraemer, H. C. y Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Sage Publications, Inc.

Kuder, G. F. y Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Mara, C., Cribbie, R., Flora, D. LaBrish, C., Mills, L. y Fiksenbaum, L. (2012). An improved model for evaluating change in randomized pretest, posttest, follow-up designs. *Methodology, 8*(3), 97-103. http/doi.org/10.1027/1614-2241/a000041

Muñiz, J. (1992). *Teoría clásica de los test*. Pirámide.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training and researchers. *Psychological Methods, 1*, 115-129. https://doi.org/10.1037/1082-989X.1.2.115

Siegel, S. (1956). *Non-parametric Statistics.* McGraw-Hill.

Tomczak, M. & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences, 21*, 19-25.

Zinbarg, R., Revelle, W., Yovel, I. y Li, W. (2005). Cronbach's, Revelle's, and Mcdonald's: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133. http/doi.org/10.1007/s11336- 003-0974-7

Aspectos de formato

American Psychological Association. (2010). *Manual de publicaciones de la American Psychological Association*. Manual Moderno.

American Psychological Association. (2020). *Publication manual.* (7.ª ed.). Manual Moderno.

Investigaciones sociales o educativas referenciadas

Alcaldía Mayor de Bogotá, Secretaría de Educación, Universidad Nacional de Colombia. Instituto de Estudios Urbanos, Bromberg Zilberstein, P., Pérez Salazar, B., Jaramillo Guerra, P. S. y Ávila Martínez, A. F. (2015). *Encuesta de clima escolar y victimización 2015.* https://repositoriosed.educacionbogota.edu.co/handle/001/459

Caballero, C., Hederich, C. y García, A. (2015). Relación entre *burnout* y *engagement* académicos con variables sociodemográficas y académicas. *Psicología desde el Caribe*, *32*(2), 254-267. http://dx.doi.org/10.14482/psdc.32.2.5742

Ministerio de Educación Nacional e Instituto Colombiano para la Evaluación de la Educación. (2017). *Estudio internacional de educación cívica y ciudadana: Informe para Colombia.* https://www.iea.nl/sites/default/files/2019-07/ICCS_2016_National_Report_Colombia.pdf

Hederich-Martínez, C. (2004). Estilo cognitivo en la dimensión de dependencia-independencia de campo. Influencias culturales e implicaciones para la educación. [Tesis doctoral, Universidad Autónoma de Barcelona]. https://www.tdx.cat/bitstream/handle/10803/4754/chm1de1.pdf

Hederich-Martínez, C., de la Portilla Maya, S. y Montoya Londoño, D. M. (2022). Características psicométricas de la escala de autoconcepto AF5 en estudiantes universitarios de la ciudad de Manizales. *Psychologia. Avances de la Disciplina*, *16*(1), 57-70. https://doi.org/10.21500/19002386.5517

Hederich-Martínez, C. y Roa-Casas, C. (2019). Análisis de publicaciones de los primeros 20 números de *Magis*. *Magis, Revista Internacional de Investigación en Educación*, *11*(23), 221-242. http://dx.doi.org/10.11144/Javeriana.m11-23.appm

Hederich-Martínez, C. (2007). *Estilo cognitivo en la dimensión de Independencia-Dependencia de Campo — Influencias culturales e implicaciones para la educación—*. [Tesis doctoral, Universidad Autónoma de Barcelona]. https://www.tdx.cat/bitstream/handle/10803/4754/chm1de1.pdf

Hederich-Martínez, C. y Camargo, A. (1999). *Estilos cognitivos en Colombia: Resultados en cinco regiones culturales colombianas.* Universidad Pedagógica Nacional, Centro de Investigaciones (CIUP)-Colciencias.

Hederich-Martínez, C., Camargo, A. y Reyes, M. E. (2004). *Ritmos cognitivos en la escuela.* Universidad Pedagógica Nacional, Centro de Investigaciones (CIUP).

Hederich, C., Camargo, A. y López, O. (2018). Motivation and use of learning strategies in students, men and women, with different level of schooling. *Journal of Psychological and Educational Research*, *26*(1), 121-146.

Londoño, R., Saénz, J., Lanziano, C., Castro, B., Ariza, V. y Aguirre, M. (2011). *Perfiles de los docentes del sector público de Bogotá.* IDEP. http://www.idep.edu.co/sites/default/files/libros/Perfiles%20de%20los%20Docentes.pdf

Mullis, I., Martin, M. y Foy, P. (2008a). *TIMSS 2007. International Mathematics Report.* TIMSS & PIRLS. International Study Center. https://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalMathematicsReport.pdf

Mullis, I., Martin, M. y Foy, P. (2008b). *TIMSS 2007. International Science Report*. TIMSS & PIRLS. International Study Center. https://timssandpirls.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalScienceReport.pdf

Pérez-Gómez, A., Lanziano, C., Reyes-Rodríguez, M., Mejía-Trujillo, J. y Cardozo-Macías, F. (2018). Perfiles asociados al consumo de alcohol en adolescentes colombianos. *Acta Colombiana de Psicología*, *21*(2), 258-281. http://www.dx.doi.org/10.14718/ACP.2018.21.2.12

Pintrich, P.; Smith, D.: Garcia, T. y Mckeachie, W. (1993). Reliability and predictive validity of the Motivated Strategies For Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801. http/doi.org/10.1177/0013164493053003024

Rincón-Camacho, L. y Hederich-Martínez, C. (2012). Escritura inicial y estilo cognitivo. *Folios*, *35*, 49-65. http://dx.doi.org/10.17227/01234870.35folios49.65

Testu, F. (1998). Chronobiologie de l'enfant, chronopsychologie scolaire et aménagements du temps. *Psychologie et Education*, *35*, 15-29.

Vega, M. L. y Hederich-Martínez, C. (2015). The impact of a Cooperative Learning Program in the academic achievement in Mathematics and Language in fourth grade students and its relations to Cognitive Style. *New Approaches in Educational Research*, *4*(2), 84-90. http://dx.doi.org/10.7821/naer.2015.7.124

Índice analítico

Sobre el autor

Christian Hederich Martínez

Matemático (Universidad Javeriana, Bogotá), magíster en Desarrollo Educativo y Social (Universidad Pedagógica Nacional, Bogotá) y doctor en Psicología (Universidad Autónoma de Barcelona). Su tesis doctoral, calificada Excelente *Cum Laude*, obtuvo el Premio Extraordinario de Doctorado de 2006. Investigador categoría Senior (Minciencias, 2022). Líder del Grupo de Investigación en Estilos Cognitivos (clasificación A1 en Minciencias, 2022). Profesor titular de la Universidad Pedagógica Nacional (Bogotá) y de la Universidad Autónoma de Manizales. Editor en jefe de la *Revista Colombiana de Educación*. Miembro del comité editorial de cinco diferentes revistas de las áreas de la psicología educativa y la educación (revistas *Hachetetepe*, *CES Psicología, Latinoamericana de Estudios Educativos, Magis* y *Suma Psicológica*). Profesor visitante en las universidades de la Frontera (Temuco, Chile), de la República (Montevideo, Uruguay) y de Alicante (Alicante, España). Autor de 12 libros y más de 80 artículos publicados en revistas indexadas.

Este libro se terminó de editar y publicar en 2023, en Bogotá, Colombia, a 269 años de la muerte del matemático británico, de origen francés, Abraham de Moivre, célebre por sus muchos aportes a la geometría analítica y a la teoría de la probabilidad; según la leyenda, predijo la fecha de su muerte a través de cálculos estadísticos.

En este texto, el profesor Christian Hederich Martínez, investigador y profesor de la Universidad Pedagógica Nacional, presenta un manual novedoso para el procesamiento y análisis cuantitativo en procesos de investigación social. Este manual recopila toda la información necesaria para el análisis y guía al investigador paso a paso desde el primer momento, cuando se elabora la base de datos, hasta la presentación final de los resultados en un artículo científico. Se agrupan así, en una única publicación, como señala Hederich, "un conjunto de elementos, trucos, procesos y conocimientos con diversos grados de estructuración, que se requieren, o que facilitan, el procesamiento y análisis cuantitativo de la información". En el texto se dan también instrucciones específicas para el uso de programas especializados en esas tareas, como el SPSS y el JASP. Así, la obra del profesor Hederich Martínez aporta una serie de herramientas para un ámbito que pocas veces es reunido y dispuesto en términos sencillos, por lo que, sin duda, se convertirá en un referente esencial para el desarrollo de procesos de investigación apoyados en el análisis cuantitativo de datos.

ISBN: 978-628-7518-91-9