Análisis de datos
en ciencias sociales y de la salud I

PROYECTO EDITORIAL:

Metodología de las Ciencias del Comportamiento y de la Salud

Directores:

Antonio Pardo Merino
Miguel Ángel Ruiz Díaz

Análisis de datos
en ciencias sociales y de la salud I

Antonio Pardo • Miguel Ángel Ruiz • Rafael San Martín

EDITORIAL SINTESIS

Consulte nuestra página web: **www.sintesis.com**
En ella encontrará el catálogo completo y comentado

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso, 34. 28015 Madrid
Teléfono 91 593 20 98
http://www.sintesis.com

ISBN:978-84-975664-7-6
Depósito Legal: M. 30.923-2009

Impreso en España - Printed in Spain

Índice de contenidos

3. Análisis descriptivo de variables categóricas

4. Análisis descriptivo de variables cuantitativas

5. Puntuaciones típicas y curva normal

9. Inferencia con una variable

10. Inferencia con dos variables categóricas

11. Inferencia con una variable categórica y una cuantitativa

12. Inferencia con dos variables cuantitativas

Presentación

Este manual de análisis de datos es el primer volumen de una serie dedicada a revisar los procedimientos estadísticos comúnmente utilizados en el entorno de las ciencias sociales y de la salud.

Por qué un nuevo manual de "análisis de datos"

La decisión de los países del viejo continente de crear el Espacio Europeo de Educación Superior ha obligado a las universidades europeas a realizar una reforma generalizada de sus planes de estudios para adaptarlos a la nueva normativa. Esta reforma, que en muchos casos ha supuesto cambios importantes, ha afectado a todas las disciplinas; y los grados agrupados bajo la denominación de "ciencias sociales" y "ciencias de la salud" no son una excepción.

Este hecho, por sí sólo, ya bastaría para justificar la presentación de un nuevo manual de análisis de datos con contenidos adaptados a los nuevos grados. Pero lo cierto es que esto sólo ha sido la excusa para elaborar una propuesta acorde con nuestra forma de entender el análisis de datos o, quizá sería más exacto decir, acorde con nuestra idea acerca de cuál es la mejor manera de iniciar a un estudiante en el análisis de datos.

Qué contenidos seleccionar

Nuestra idea de cómo hacer las cosas afecta tanto a los contenidos seleccionados como a la forma de presentarlos. En la selección de los contenidos se ha tenido en cuenta, por un lado, la presencia cada vez más extendida de programas informáticos y de ordenadores donde poder utilizarlos; y, por otro, el hecho de que, por lo general, los estudiantes, profesores e investigadores que se mueven en el ámbito de las ciencias sociales y de la salud, ni son estadísticos ni pretenden serlo.

En primer lugar, el poder contar con programas informáticos capaces de aplicar cualquier procedimiento estadístico con suma facilidad y con el mínimo esfuerzo ha convertido en obsoletos algunos procedimientos a los que antes se les dedicaba bastate atención (por ejemplo, todo lo relativo a la agrupación de variables en intervalos o a los métodos abreviados de cálculo); dejar de lado estos procedimientos ha permitido liberar espacio para incluir otros nuevos. Pero además, ya no es necesario invertir tiempo haciendo a mano cálculos que no contribuyen en absoluto a entender el significado de lo que se está haciendo (como, por ejemplo, aplicar la fórmula clásica del coeficiente de correlación de Pearson para cuantificar el grado de relación lineal entre dos variables), lo cual contribuye de forma significativa a no tener que desviar la atención de lo realmente importante, que, en nuestra opinión, no es precisamente

realizar cálculos, sino aprender a elegir el procedimiento apropiado en cada caso y a interpretar correctamente los resultados que ofrece. Aunque todos los procedimientos se presentan con suficiente detalle como para poder aplicarlos a mano, de todos ellos se explica también cómo aplicarlos con un programa informático.

En segundo lugar, no se presta atención detallada a algunos contenidos que, por ser fundamento matemático de los procedimientos estadísticos, es habitual encontrarlos en la mayoría de los manuales sobre análisis de datos y estadística aplicada. En este sentido, no se ofrece ningún capítulo dedicado a contenidos que suelen recibir bastante atención en otros manuales: la teoría de la probabilidad, que suele merecer al menos un capítulo, se resume en un apartado; las distribuciones de probabilidad, que suelen tratarse como un bloque en uno o dos capítulos, como un listado de distribuciones independientes de lo demás, aquí se presentan como parte complementaria de otros procedimientos y sólo se presta atención a las más importantes; y todo lo relacionado con el estudio pormenorizado de la variables aleatorias (valores esperados y momentos, funciones de probabilidad, etc.) se ha reducido a la mínima expresión. Un profesional de las ciencias sociales y de la salud no es un estadístico; y, muy probablemente, tampoco pretende serlo; consecuentemente, no necesita ser un experto en los fundamentos matemáticos de las herramientas estadísticas. Creemos que el énfasis hay que colocarlo, más bien, en conocer la utilidad de los procedimientos disponibles y en saber elegirlos, aplicarlos e interpretarlos correctamente.

Cómo presentar los contenidos

En nuestra idea acerca de cuál es la mejor manera de iniciar a un estudiante en el análisis de datos también desempeña un papel importante la forma de presentar los contenidos. Y esto afecta tanto a la ordenación de los mismos como a la forma de exponerlos.

En lo relativo a la ordenación de los contenidos, nuestra opción, aunque poco convencional, nos ha parecido que era la mejor apuesta. Por lo general, los manuales de introducción al análisis de datos presentan una primera parte con herramientas descriptivas para una variable (distribuciones de frecuencias, gráficos, estadísticos de posición, dispersión y forma, etc.) y para dos o más variables (distribuciones conjuntas, relación lineal, regresión lineal), dedicando una segunda parte a la lógica de la inferencia estadística y a algunas herramientas inferenciales concretas para una y dos variables (con atención ocasional al estudio de más de dos variables). En nuestra propuesta, las herramientas descriptivas se explican referidas a una sola variable; después se introducen los conceptos inferenciales y, a partir de ahí, el análisis de dos o más variables se realiza mezclando las herramientas descriptivas con las inferenciales.

Varias razones nos han hecho optar por este formato. La primera es más bien de tipo, podríamos decir, *profesional*. Cuando un analista de datos se enfrenta con un archivo de datos, la primera tarea que aborda es la de intentar formarse una idea lo más exacta posible acerca de las características de los datos. Esta tarea se lleva a cabo aplicando, sin ideas preconcebidas, herramientas descriptivas y exploratorias a cada una de las variables individualmente consideradas intentando identificar tanto regularidades como anomalías. En una segunda fase, el analista mezcla variables para obtener nueva información sobre las características de los datos. Pero cuando se estudian dos o más variables simultáneamente es porque interesa, no sólo describirlas, sino compararlas o relacionarlas de acuerdo con un plan preconcebido (no se mezcla todo con todo); y para esto no suele ser suficiente aplicar herramientas descriptivas;

son las herramientas inferenciales las que ayudan a detectar la posible presencia de diferencias o relaciones. De ahí que, en el estudio de más de una variable, nos parezca conveniente estudiar simultáneamente las herramientas descriptivas y las inferenciales.

La segunda razón es de tipo *docente*. Según nuestra experiencia, los conceptos inferenciales y la lógica en la que se basan son los contenidos que más cuesta asimilar a quienes se acercan por primera vez al análisis de datos. Para facilitar la comprensión de estos contenidos nos ha parecido buena idea darles dos repasos: uno más básico e intuitivo (en este primer volumen) y otro algo más profundo y fundamentado (en el segundo volumen). De ahí que en este primer volumen hayamos decidido incluir varios capítulos dedicados a la lógica de la inferencia estadística y a la aplicación de algunas herramientas inferenciales concretas.

Por lo que se refiere a la forma de exponer los contenidos, se ha intentado prestar más atención a los aspectos prácticos o aplicados que a los teóricos o formales, pero sin descuidar estos últimos. Aunque este manual va dirigido, principalmente, a estudiantes de disciplinas del ámbito de las ciencias sociales y de la salud, no se trata de un material diseñado exclusivamente para ellos. También pretende servir de ayuda a los profesores de análisis de datos y a los investigadores. Creemos que ambos pueden encontrar, en éste y en los siguientes volúmenes, las respuestas a muchas de las preguntas que se hacen en su trabajo cotidiano.

Y todo ello sin olvidar que, en los tiempos que corren, no tiene sentido analizar datos sin el apoyo de un programa informático. Ahora bien, conviene tener muy presente que, aunque las herramientas informáticas pueden realizar cálculos con suma facilidad, todavía no están capacitadas para tomar algunas decisiones. Un programa informático no sabe si la estrategia de recogida de datos utilizada es la correcta, o si las mediciones aplicadas son apropiadas; tampoco decide qué prueba estadística conviene aplicar en cada caso, ni interpreta los resultados del análisis. Los programas informáticos todavía no permiten prescindir del analista de datos. Es él, el analista, quien debe mantener el control de todo el proceso. El éxito de un análisis depende de él y no del programa informático. El hecho de que sea posible ejecutar las técnicas de análisis más complejas con la simple acción de pulsar un botón sólo significa que es necesario haber atado bien todos los cabos del proceso (diseño, medida, análisis, etc.) antes de pulsar ese botón.

No podemos dejar pasar la oportunidad que nos brinda esta presentación para agradecer a nuestro compañero Ludgerio Espinosa, y a muchos de nuestros alumnos y a no pocos lectores de nuestros trabajos previos, las permanentes sugerencias hechas para mejorar nuestras explicaciones y la ayuda prestada en la caza de erratas. Los errores y deficiencias que todavía permanezcan son, sin embargo, atribuibles sólo a nosotros.

Antonio Pardo
Miguel Ángel Ruiz
Rafael San Martín