Introducción a la Estadística

PID_00247916

Ester Bernadó

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Índice

Introducción

Son muchos los expertos que etiquetan nuestro tiempo como la *era de la información*, también denominada la *era de la información y las comunicaciones* o la *era digital*. Y probablemente estaremos muy de acuerdo, pues el impacto de la información y las tecnologías de la información en nuestras vidas no pasa desapercibido. Son múltiples las actividades diarias que están influidas por las tecnologías (por prudencia, no usaremos el término «determinadas»). Entre ellas, el móvil y sus numerosas aplicaciones, las redes sociales, el correo electrónico o las reuniones virtuales. Ya no podemos concebir formas de trabajar, de desplazarnos, de comunicarnos, de establecer relaciones… sin el uso de estas tecnologías. Alguien se preguntó una vez: «¿Dónde se buscaba la información antes de Google?».

De la mano de la era de la información, se está produciendo otra revolución, subyacente y a menudo escondida a nuestros quehaceres diarios. Esta es la *era de los datos* (Mayer-Schönberger y Cukier, 2013). Y es que todo aquello que realizamos con las tecnologías, como enviar un correo, publicar un mensaje en una red social, usar un navegador, realizar una transacción bancaria o realizar una compra en el supermercado queda registrado. Dicho de esta forma, quizás el estudiante establecerá una analogía con la distopía de George Orwell narrada en su novela *1984*, donde se describe una sociedad permanentemente controlada por el Gran Hermano, el cual vigila la población mediante «telepantallas». No es la intención de este módulo adentrarnos en un debate como este, sino despertar el interés por los datos y por las múltiples oportunidades que su análisis ofrece para mejorar la calidad de vida de las personas, el beneficio económico de un negocio, los diagnósticos médicos o la comprensión de nosotros mismos, por citar tan solo algunos ejemplos. Explican Mayer-Schönberger y Cukier (2013) que se puede predecir en tiempo real el avance de la gripe en Estados Unidos por las búsquedas en Google, a diferencia de las predicciones con datos epidemiológicos que solo son capaces de detectar un brote de gripe con dos semanas de retraso.

Más allá de los datos personales registrados, existe un gran volumen de datos que se capturan a partir de distintos dispositivos. Bajo el concepto de *internet de las cosas* se incluye cualquier dispositivo físico o virtual que es capaz de recolectar datos y compartirlos en red, como edificios, vehículos, y en general cualquier dispositivo electrónico, sensores y *software*. Con ello se abre la puerta a un mundo de aplicaciones infinitas, como los edificios inteligentes, la optimización del transporte público, la gestión inteligente de la red eléctrica, el control y gestión preventiva de tráfico o las ciudades inteligentes.

Son tres los elementos básicos de esta *era de los datos*:

- Los datos en sí, o bancos de datos, donde se almacena la información.

- Los algoritmos o *software* computacional capaz de extraer información de interés de estos datos.

- Las aplicaciones, o aquello que se hace a partir de estos análisis.

El gran avance tecnológico ha propiciado el almacenaje de enormes volúmenes de datos y una gran capacidad computacional para realizar análisis automáticos y eficientes, generando un gran interés en las infinitas posibilidades de aplicación.

El interés por los datos no nace con la era de la información. Se origina con la curiosidad humana para comprender el mundo en que vivimos, predecirlo y dominarlo. Matemáticos, astrónomos, físicos, antropólogos, biólogos e historiadores han recolectado datos, realizado análisis y extraído conclusiones que han permitido el progreso de la ciencia y la sociedad. La historia de Johannes Kepler y Tycho Brahe es un buen ejemplo. Tycho Brahe (1546-1601) fue un noble danés que dedicó gran parte de su vida a realizar observaciones precisas del universo, antes de la invención del telescopio. Observó y anotó las posiciones relativas de los planetas de forma mucho más precisa que las realizadas en la época. Johannes Kepler (1571-1630) fue un matemático y astrónomo que quería demostrar la estructura geométrica perfecta y divina de las órbitas de los planetas conocidos hasta entonces. Pero no disponía de datos suficientes para demostrar la órbita circular y concéntrica de los planetas de su teoría. El encuentro entre Brahe y Kepler permitió la unión de unas observaciones sistemáticas y precisas realizadas durante 35 años, con la pasión por extraer un modelo de estos datos. A la muerte de Brahe en 1601, Kepler pudo disponer de sus observaciones. Kepler se esforzó por encontrar un modelo matemático que encajara en la teoría, pero no conseguía explicar las observaciones de Brahe. Obstinado y perseverante, no quiso despreciar un error de 8 minutos de arco en una órbita ni dar las observaciones por erróneas. Así que finalmente, después de varias iteraciones, renunció a la órbita circular y probó con una órbita elíptica, con la que llegó a enunciar las conocidas tres leyes de Kepler.

En el corazón de los análisis de datos se sitúan las matemáticas y en especial la estadística. El término actual *estadística*, introducido originalmente por el alemán Gottfried Achenwall como *statistik* en 1749, proviene del latín *statisticum collegium* («consejo de estado») y de su derivado italiano *statista* («hombre de estado», «político»). Se refería al análisis de datos de estado, especialmente orientado a la recolección sistemática de datos demográficos y económicos por parte de los estados. De hecho, esta orientación se originó ya en las civilizaciones antiguas como la egipcia (3.050 años a. C.), la china (2.200 años a. C.) y la antigua Roma, donde se recolectaban censos demográficos para la pla-

nificación de la agricultura y la economía (como la captación de impuestos). La estadística es una disciplina en constante evolución, que se ha enriquecido con distintas aproximaciones, técnicas y procedimientos.

Uno de los censos más famosos de la historia romana y quizás de toda la historia occidental fue precisamente realizado en el año 0. La Biblia narra cómo María y José viajaban hacia Belén para formar parte de los datos censales romanos. En el camino, María dio a luz a Jesús. Así es como el censo romano y la incipiente estadística fueron cruciales para el nacimiento de la Cristiandad.

Actualmente, la estadística sienta las bases del análisis de datos, tal como veremos a lo largo de este módulo. En primer lugar, se describe la estadística como la ciencia de los datos, definición que se matiza en los apartados siguientes mediante las principales dimensiones de aplicación de la estadística y la distinción entre estadística descriptiva e inferencial. Puesto que la estadística es un enfoque necesario para el análisis de datos pero no exclusivo, se introduce el concepto de minería de datos y *machine learning* («aprendizaje automático») por su relevancia y actualidad dentro del análisis de datos. Es necesario definir y relacionar la estadística y la minería de datos, puesto que son disciplinas con un elevado grado de relación, tanto en sus aplicaciones como en el uso de algunos métodos, a la vez que provienen de enfoques relativamente distintos. El objetivo del apartado es centrar las bases para la comprensión general de la estadística en el marco del análisis de datos y preparar al estudiante para la profundización en sus teorías y métodos.

1. ¿Qué es la estadística?

Históricamente, la estadística ha sido la ciencia de recolección, análisis, interpretación, presentación y organización de los datos. Algunos expertos simplemente definen estadística como la *ciencia de los datos*. ¿Pero qué son los datos? Haciendo una simplificación, los datos son números en contexto. Y es que la estadística va más allá de realizar cálculos sobre los datos. Consiste en interpretar estos cálculos en el contexto en que se producen con el objetivo de describir la información inherente, predecir comportamientos futuros o tomar decisiones. Moore, McCabe y Craig (2012) lo ejemplifican de la siguiente manera:

> «Calcular la media y desviación del peso al nacer de 1000 niños es simple aritmética, interpretar además su significado es estadística. La estadística se fundamenta en la teoría de probabilidades, que describe el modelo matemático inherente a los fenómenos aleatorios. La estadística también involucra el juicio, que es necesario para aplicar el procedimiento adecuado en función de las condiciones del problema.»
>
> Moore, McCabe y Craig (2012)

Profundicemos un poco más en los datos. A menudo se confunde el término datos con información. En efecto, los datos son una fuente de información, pero esta información no es obvia, ni inmediata. Imaginemos por ejemplo un fichero de datos sobre los clientes de un banco a los que se les concede un crédito, tal como muestra la tabla 1. Supongamos que disponemos de 2.000 casos (registros) como los descritos. ¿Qué tipo de información proporciona la tabla? A simple vista, es un listado de datos, que necesita ser analizado, interpretado y presentado en una forma que aporte información útil.

Tabla 1. Datos de solicitud de préstamos de una entidad bancaria

ID	Nombre	Edad	Sexo	Estado civil	Número de hijos	Nivel salarial	Crédito solicitado	Préstamo hipotecario	Motivo de préstamo
567	José Pérez	46	H	C	2	3	20.000	65.000	Vehículo
765	María Sol	34	M	S	0	2	5.000	0	Estudios
965	Simón Martín	52	H	C	2	4	350.000	0	Reformas

La estadística permite extraer información de los datos. Por ejemplo, algunas de las preguntas que pueden realizarse son:

- ¿Cuántos hijos en promedio tienen las personas que piden créditos?

- ¿Cuál es el estado civil más frecuente entre las personas que piden créditos?
- ¿La edad media de las personas que solicitan un crédito es 35 años?
- ¿Las mujeres piden créditos de valor superior al de los hombres?
- ¿Existe relación entre el nivel de ingresos y la cantidad solicitada de crédito?
- ¿El importe del crédito solicitado es diferente según el motivo de préstamo?

En definitiva, los datos en sí no aportan información. Es mediante el uso de técnicas de análisis de datos que podemos plantear preguntas sobre los datos y conocer las correspondientes respuestas. Como veremos más adelante, estas preguntas se formulan en forma de hipótesis, que se someten a prueba mediante el uso de las técnicas estadísticas adecuadas, las cuales llevan al rechazo o aceptación de las hipótesis.

Por ejemplo, se puede formular la hipótesis siguiente: las mujeres piden créditos de valor superior al de los hombres.

A partir de la muestra disponible, se pone a prueba la hipótesis y el resultado será el rechazo o la aceptación de la hipótesis con un determinado nivel de confianza (el cual se expresa como un porcentaje, como por ejemplo con un 95 % de nivel de confianza).

Así es como, a grandes rasgos y de forma muy simplificada, la estadística realiza una extracción de información de los datos. De hecho, hay un conjunto concreto de preguntas (o hipótesis) que se pueden realizar sobre los datos. Conociendo estos tipos de preguntas, en qué casos deben aplicarse (las condiciones de aplicación) y la interpretación de las respuestas, tenemos un espectro amplísimo de extracción de información sobre los datos.

Decíamos en la definición que la estadística también interviene en la recolección de los datos. A menudo, los datos de los que se dispone para el análisis están previamente guardados en bases de datos, y son los datos que se usan por su disponibilidad. En otras ocasiones, es posible diseñar la recolección de datos, y es en estos casos donde la estadística puede orientar en la recolección de una muestra representativa de la población de interés. Asimismo, la estadística también aporta herramientas adicionales de interpretación de la información mediante el uso de gráficos y visualizaciones.

El objetivo de un curso de estadística sería el de dotar al estudiante de las herramientas de análisis estadístico, con las cuales es capaz de hacer las preguntas adecuadas, recolectar los datos necesarios, aplicar las técnicas adecuadas a ca-

da situación y extraer las interpretaciones adecuadas, usando visualizaciones si es necesario y sabiendo juzgar además los errores y/o márgenes de confianza de las conclusiones obtenidas.

2. Aplicaciones de la estadística

La estadística está presente en multitud de dominios: las ciencias sociales, los negocios, el diagnóstico y tratamiento médico, el control de calidad de los productos, la predicción meteorológica, la educación, la investigación, y un largo etcétera. Allí donde hay datos, allí está la estadística. Newbold (1997) agrupa en 6 tipos las aplicaciones de la estadística. Comentábamos antes que el tipo de preguntas o análisis estadísticos que se pueden realizar sobre los datos era finito. De forma análoga, Newbold agrupa estos tipos de preguntas o análisis en seis:

- Extracción de conclusiones de variables numéricas
- Manejo de la incertidumbre
- Análisis de relaciones
- Muestreo
- Predicción
- Toma de decisiones con incertidumbre

En los apartados siguientes se puede ver una breve descripción de cada tipo junto con varios ejemplos.

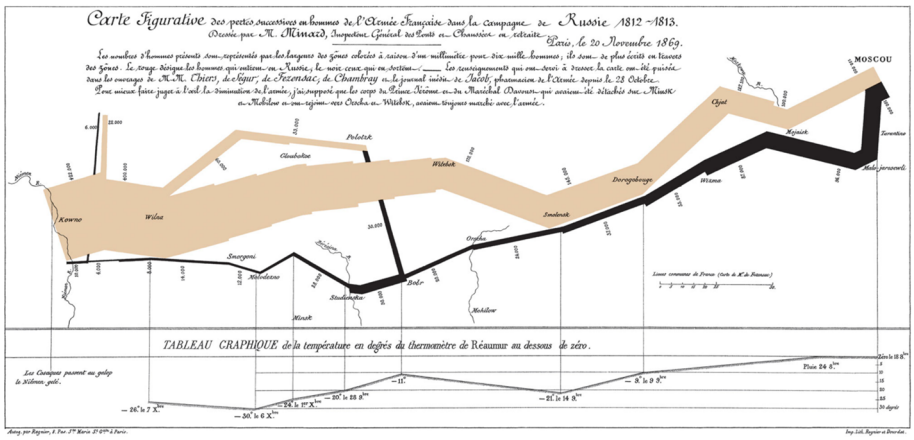
2.1. Extracción de conclusiones de variables numéricas

Una parte importante de la estadística trata con datos numéricos recolectados en forma de lista de datos y que a menudo suele tener un volumen grande. La función de la estadística es extraer y sintetizar las características fundamentales de esta lista de datos numéricos.

El interés por los datos numéricos y estadísticos básicos es muy antiguo. En la Antigüedad, los babilonios, los chinos, los egipcios, los griegos y los romanos hicieron censos de población. Con ello, pretendían estimar cuánto dinero podía recogerse con los impuestos, calcular cuántos soldados se podían reclutar para un ejército o cuánta comida haría falta (Rooney, 2009).

En 1662, John Graunt, considerado como el padre de la demografía y de la estadística, publicó unas estadísticas sobre la mortalidad en Londres. Mostró el número de muertes por enfermedad como un intento de aliviar la ansiedad de la población en relación a determinadas enfermedades. También fue el primero en documentar que nacían más niños que niñas. Proporcionó una estimación de la población de Londres, demostrando su rápido crecimiento. Mostró que la disminución de la población en una plaga se compensaba con un rápido crecimiento debido al aumento de los nacimientos (Rothman, 1996). Otra notable contribución histórica es el gráfico de Charles Minard sobre la campaña de Napoleón en Rusia en 1812. El siguiente gráfico muestra visual-

mente el tamaño del ejército en la ida y en la vuelta a lo largo de los kilómetros recorridos y la temperatura. Solo 4 de cada 100 soldados regresaron de la incursión napoleónica (Rooney, 2009).



Las medidas más habituales para resumir la información numérica son las de media, moda o valor más frecuente, las medidas de dispersión y los rangos intercuartílicos. Se acompañan a menudo de gráficos típicos como los histogramas o diagramas de caja.

2.2. Manejo de la incertidumbre

Tomeo y Uña (2003) definen la estadística como:

> «La ciencia que utiliza los números para el estudio de las leyes que dependen del azar, tratando de descubrir mediante el razonamiento inductivo la causa general a que obedece el modelo particular analizado.»
>
> Tomeo y Uña (2003)

Si prestamos atención a la primera parte de la definición, el estudio del azar o los fenómenos aleatorios se refiere al estudio de aquellos fenómenos cuyo resultado es distinto, aunque se produzca en las mismas condiciones, en contraposición a los fenómenos *deterministas*, que producen el mismo resultado ante las mismas condiciones.

Estudio del azar

El lanzamiento de una moneda o un dado no da siempre el mismo resultado.

En estos casos, en lugar de certeza, hablamos de probabilidades: la probabilidad de que salga cara al lanzar una moneda, o un 6 al lanzar un dado. Newbold se refiere a la estadística como *la ciencia de la incertidumbre*.

Galileo afirmaba que la ley del azar era una muestra de la incapacidad humana. Según Galileo, lo que hace que el lanzamiento de una moneda genere un resultado impredecible es que no se ha lanzado la moneda en idénticas condiciones (Tomeo y Uña, 2003).

2.3. Análisis de relaciones

El análisis de relaciones entre las variables de un problema es una de las cuestiones más frecuentes que aborda la estadística.

Un ejemplo ilustrativo se presentó a raíz de la pregunta sobre si existía relación entre el cáncer de mama y la toma de anticonceptivos orales (Armitage, Berry y Matthews, 2002, pág. 4). Para dar respuesta a ello, en 1990 la Organización Mundial de la Salud realizó un estudio en 12 centros de 10 países distintos con el título *Collaborative Study of Neoplasia and Steroid Contraceptives*. En cada hospital se escogieron casos de mujeres con cáncer de mama que cumplían determinados requisitos de edad y residenciales. Otra muestra de control se tomó a partir de mujeres que visitaron el mismo hospital, con criterios de edad y residenciales similares, y que no tomaban anticonceptivos orales. El estudio incluyó 2.116 casos y 13.072 casos de control. Se estudiaron otras variables como la edad, la edad del primer hijo, el índice socio-económico y el historial familiar de cáncer de mama. El riesgo de cáncer de mama para usuarias de anticonceptivos orales fue identificado como 1,15 mayor que las que no tomaban anticonceptivos, lo cual demostró ser una asociación muy débil en comparación con otros factores más influyentes.

El análisis de relaciones permite estudiar si las variables identificadas de un determinado ámbito o dominio varían conjuntamente. Con ello, se identifican variables que están asociadas entre sí.

Análisis de relaciones

El consumo eléctrico de los hogares es superior si el día es muy caluroso o frío; la contaminación atmosférica de una ciudad está relacionada con el número de coches que circulan y con las condiciones atmosféricas.

Habitualmente, se usa el término específico *análisis de correlaciones* para denominar este tipo de estudios. Un caso particular del mismo es el *análisis de correlaciones lineales*, el cual mide el grado de dependencia lineal entre un conjunto de variables. Asimismo, conviene realizar una distinción entre correlación y causalidad. Si dos variables se demuestran correlacionadas, no significa que exista una relación de causalidad, es decir, que una de ellas sea la causa de la variación de la otra. Tendemos a atribuir demasiado fácilmente efectos de causalidad entre variables correlacionadas. El caso es tan frecuente que la frase «*correlation doesn't imply causation*» (la correlación no implica causalidad) tiene incluso una entrada en Wikipedia y se escriben libros con curiosas correlacio-

nes, como que el número de suicidios por ahogamiento o estrangulamiento está correlacionado con el gasto del Gobierno de Estados Unidos en ciencia, espacio y tecnología (Vigen, 2015).

Además del análisis de correlaciones, otra forma de estudiar las relaciones entre variables es a través de la construcción de **modelos de regresión**.

Los modelos de regresión tratan de extraer una función matemática que relacione las variables entre sí. El modelo más simple es el modelo de regresión lineal. El análisis de relaciones se acompaña asimismo de gráficos ilustrativos como los gráficos de dispersión.

Cuando se hacen este tipo de análisis con finalidad predictiva es habitual denominar a estas relaciones de «causa-efecto».

2.4. Muestreo

En el estudio mencionado sobre la posible relación entre cáncer de mama y uso de anticonceptivos, se deseaba demostrar la relación entre estas variables en la población. Sin embargo, como no era posible abarcar toda la población, se realizó el estudio sobre una muestra con el fin de generalizar los resultados a toda la población.

Un ejemplo típico de muestreo es el de los sondeos de opinión en días previos a unas elecciones. Los periódicos suelen mostrar los resultados de la intención de voto de la población, resultados que evidentemente se han generalizado a partir de una muestra de la población. En estos casos, cuando se escoge una muestra de la población, el objetivo no es describir la muestra en sí misma, sino generalizar las conclusiones extraídas en la muestra a toda la población.

Los métodos de muestreo que usan aleatorización planificada se denominan *muestreos probabilísticos*. La variante más básica es el *muestreo aleatorio simple*. En este, todos los individuos de la población deben tener la misma probabilidad de pertenecer a la muestra. Imaginemos que toda la población de interés aparece en un listín telefónico. Cada persona tendría asignado un número entre 1 y N, donde N es el tamaño de la población. El muestreo aleatorio simple consistiría en seleccionar k individuos al azar, escogiendo cada vez uno de los individuos entre 1 y N de forma aleatoria. De esta forma, cada individuo tiene la misma probabilidad de pertenecer a la muestra.

Cuando este tipo de procedimientos no son posibles (habitualmente no se dispone de una lista de la población de interés), existen otros esquemas de muestreo como el muestreo sistemático, el muestreo estratificado y el muestreo por conglomerados.

Para ejemplificar uno de ellos, el *muestreo estratificado* (Scheaffer, 1999) consiste en separar la población en grupos disyuntos, denominados estratos, y seleccionar una muestra aleatoria simple de cada estrato. El tamaño de la muestra es también relevante, siendo deseable trabajar con muestras grandes; aunque suele ser más costoso, lento y/o difícilmente accesible. El método de muestreo determina en gran medida la representatividad de la muestra y, por tanto, las generalizaciones que se extraigan pueden ser más o menos precisas.

En 1934 el estadístico y matemático Jerzy Neyman diseñó el primer método de muestreo estratificado. Dos años más tarde, el sondeo electoral Gallup predijo la victoria de Roosevelt en las elecciones de presidencia de Estados Unidos usando el muestreo estratificado. Lo más curioso del caso es que la predicción del muestreo estratificado predijo correctamente la victoria, a diferencia de otro sondeo realizado con una muestra mucho más amplia de la población que predecía la victoria del oponente. Desde entonces el sondeo electoral de George Gallup usa este tipo de muestreos (Scheaffer, 1999).

2.5. Predicción

A menudo los datos con los que trabaja la estadística se hallan previamente guardados en bases de datos y representan una fotografía estática del dominio a estudiar. El ejemplo sobre la concesión de créditos a clientes sería uno de estos casos. Otros ejemplos podrían ser el rendimiento de los estudiantes de un curso, el diagnóstico y tratamiento de los pacientes de una consulta médica, los productos con más ventas de un determinado negocio, etcétera. Los análisis típicos aplicables pueden ser de tipo descriptivo, como extraer valores numéricos de la muestra y análisis de relaciones, tal como se ha ejemplificado anteriormente.

La estadística también se ocupa de otro tipo de datos que están asociados a un componente temporal. En estos casos, los datos siguen una secuencia temporal, que se ha recolectado a lo largo de un periodo de tiempo y el objetivo es predecir la evolución de los datos en el futuro. Por ejemplo, la temperatura máxima y mínima de cada día en Barcelona a lo largo de los últimos diez años. Otro ejemplo clásico es la evolución de la bolsa. El estudio del comportamiento de los datos en el futuro se aborda desde los modelos de *series temporales*. Por una parte, el modelado de una serie temporal permite ajustar un modelo a los datos pasados y descomponer el mismo en sus componentes principales (tendencia, estacionalidad, aleatoriedad). A partir del modelado de la serie temporal, puede predecirse la evolución en el futuro.

En 1909, un estadístico de Bell Telephone predijo el número de operadoras requeridas en centrales telefónicas para atender la creciente demanda de llamadas telefónicas. La predicción de la demanda futura mostró que todas las mujeres americanas entre 17 y 60 años tendrían que trabajar como operadoras hacia el año 1930 para satisfacer el volumen esperado de llamadas. La predicción aceleró la invención del primer conmutador automático, que fue diseñado y puesto en servicio por Bell dos años más tarde de la predicción (Drucker, 2006).

2.6. Toma de decisiones con incertidumbre

En multitud de situaciones, se necesitan tomar decisiones entre un conjunto de opciones alternativas, sin conocer de antemano las consecuencias de estas alternativas. Existe un elevado grado de incertidumbre sobre cómo será el comportamiento futuro, debido al efecto incierto de determinados factores.

Por ejemplo, un empresario debe decidir si invertir su presupuesto en la línea de producción del producto A o del producto B, sabiendo que la inversión repercutirá en un aumento de producción. Sin embargo, desconoce con exactitud la demanda futura de estos productos, con lo cual no puede asegurar qué inversión resultará más rentable. El empresario puede imaginar diferentes escenarios posibles y a partir de los mismos, aplicar un criterio que ayude a minimizar su riesgo o maximizar sus ganancias. Para ello, existen distintos métodos, desde los basados en cálculos de maximización de las ganancias o minimización de pérdidas, hasta métodos que incorporan cálculos probabilísticos.

3. Estadística descriptiva y estadística inferencial

3.1. Estadística descriptiva

La estadística descriptiva se refiere al conjunto de técnicas para recolectar y presentar datos numéricos (Armitage, Berry y Matthews, 2002, pág. 2). Según Tomeo y Uña (2003), la estadística descriptiva trata de la descripción numérica de conjuntos, siendo particularmente útil cuando estos son de muchos elementos, valorando matemáticamente y analizando el colectivo representado por el conjunto sin pretender obtener conclusiones más generales, lo que es objeto de la estadística inferencial. En definitiva, la estadística descriptiva trata de resumir los datos de una muestra, en lugar de extraer conclusiones sobre la población en general.

Básicamente, la estadística descriptiva realiza tres tipos de funciones:

- En primer lugar, se estudia la distribución de la variable o variables de interés (qué valores toma y cómo se distribuyen estos valores).

- En segundo lugar, calcula resúmenes numéricos de estos valores como las medidas de tendencia central (la media y la mediana) y la dispersión (desviación estándar).

- Por último, se encarga de visualizar gráficamente esta información, de forma que se puede comprender rápidamente cómo es la distribución de los datos, en relación a su dispersión y simetría.

Estadística descriptiva

Se disponen de datos del peso al nacer de 500 niños nacidos en el año 1970 en un hospital de Barcelona. La tendencia central de estos datos arrojaría que la media de peso al nacer es de 3,54 kg y la mediana de 3,30 kg. El estudio de la dispersión podría resultar en una desviación estándar de 1,30 kg, que correspondería a la media de las diferencias al cuadrado respecto el peso medio. La extracción de estos valores se realiza para la muestra específicamente, sin intención de generalizar las conclusiones a toda la población de niños nacidos en 1970 en Barcelona. Otros análisis descriptivos se pueden realizar mediante visualizaciones gráficas, donde se aportaría información adicional de cómo se distribuyen estos pesos.

3.2. Estadística inferencial

La estadística inferencial va más allá de la estadística descriptiva y consiste en la obtención de resultados que generalizan los comportamientos observados en los datos y que permitan extraer conclusiones de carácter más general sobre la población (Alea *et al.*,1999). Según la definición de Gibergans, Gil y Rovira (2009), la estadística inferencial se basa en la extracción de conclusiones so-

bre una población a partir de una muestra (un subconjunto de los individuos de la población) y precisar con qué márgenes de confianza son válidas estas afirmaciones.

Los métodos de inferencia estadística se dividen principalmente en dos:

- Métodos de estimación de parámetros
- Métodos de contrastes de hipótesis

La estimación de parámetros consiste en fijar valores concretos a los parámetros que caracterizan la distribución de probabilidad de la población. El contraste de hipótesis permite validar hipótesis estadísticas que hacen referencia al valor de un parámetro poblacional (el valor esperado, la variancia, la proporción, etc.) o la relación que existe entre parámetros análogos de dos poblaciones.

En términos generales, el contraste de hipótesis permite decidir si la evidencia empírica aportada por la muestra es o no compatible con la hipótesis referida a la población sobre la cual se intenta generalizar (Alea *et al.*, 1999). Este tipo de análisis se presenta con frecuencia en múltiples dominios: en medicina, para comparar la eficacia de dos tratamientos distintos; en agricultura, para conocer el mejor fertilizante; en educación, para comparar el mejor método de estudio; en marketing, para conocer la campaña con más adquisición de clientes; etc.

En el ejemplo citado anteriormente, la muestra de estudio compuesta por 500 niños nacidos en un hospital de Barcelona en 1970 podría servir de base para inferir parámetros sobre la población. Por ejemplo, se podría deducir que el peso de los niños y niñas al nacer en la Barcelona del 1970 está comprendido entre 3,43 kg y 3,66 kg con un nivel de confianza del 95 %. Mediante contrastes de hipótesis, se puede validar si la media de los niños al nacer es igual o superior a un determinado valor (por ejemplo, 3,5 kg), o bien si la media del peso al nacer de los niños es superior al de las niñas. También se podrían comparar los parámetros de dos poblaciones de interés (recogidas en 1970 y en 2010 en Barcelona) y preguntarnos si los niños y niñas nacidos en 2010 tienen un peso superior a los niños y niñas nacidos en 1970. Como ya hemos mencionado, la representatividad de las muestras es crucial para que las conclusiones extraídas sean precisas.

4. Estadística frente a minería de datos

A partir de la «moda de los datos», se ha despertado un creciente interés en la estadística y otras disciplinas, como el *machine learning* (traducido como «aprendizaje automático») y el *data mining* («minería de datos»); a la vez que se ha reetiquetado la estadística con nombres como analítica o ciencia de los datos y han emergido otros como *big data* (Mayer-Schönberger y Cukier, 2013). En este nuevo panorama, es difícil distinguir qué dominio pertenece a cada disciplina. La línea que separa estadística y minería de datos es muy fina.

Tradicionalmente, la estadística proviene del interés muy incipiente de las civilizaciones antiguas por realizar censos de población, fundamentándose en las matemáticas y la teoría de las probabilidades. La minería de datos surge de un enfoque computacional, a partir del uso de ordenadores para almacenar bases de datos y del desarrollo de algoritmos computacionales que permitan incorporar inteligencia artificial y/o aprendizaje artificial para extraer información de estos datos, aprovechando la gran potencia de cálculo, y alimentar con ello procesos de decisión.

Estadística y minería de datos tienen muchos puntos de encuentro y por eso es difícil marcar una línea divisoria. Sin embargo, puede ser pedagógico (aunque probablemente reduccionista) contrastar ambas disciplinas, para que el estudiante pueda hacerse una imagen preliminar que podrá ir enriqueciendo a medida que se adentre en los detalles de cada disciplina. Veamos pues qué es la minería de datos en comparación con la estadística.

Según Han y Kamber (2001), el término *minería de datos* se refiere a la extracción de conocimiento de grandes cantidades de datos. Principalmente, la disciplina se ha centrado en la extracción de patrones de bases de datos. Un término similar ya un poco en desuso es el de KDD (*Knowledge Discovery in Databases*) el cual engloba todo el proceso de análisis desde la preparación de datos hasta su interpretación (Han y Kamber, 2001):

1) Limpieza de datos
2) Integración de datos
3) Selección de datos
4) Transformación de datos
5) Extracción de conocimiento (minería de datos)
6) Evaluación de patrones
7) Presentación del conocimiento

Pyle (1999) dice que históricamente el análisis estadístico se ha orientado a la verificación y validación de hipótesis; una aproximación que se ha visto influida por la ciencia. Se establece una hipótesis, se recogen las evidencias y las evidencias se confrontan con la hipótesis para ver si esta puede ser rechazada o no. Según Pyle, la minería de datos le da «la vuelta a la tortilla». En lugar de establecer las hipótesis a testear, el analista toma un conjunto de datos y pregunta: «¿Cuáles son las hipótesis que este conjunto de datos soporta?». Añade Pyle otro elemento a tener en cuenta: los grandes volúmenes de datos. La estadística usa métodos en los que es necesario imaginar potenciales relaciones entre los datos, que son posteriormente validadas con los métodos de análisis apropiados. Pero los grandes volúmenes de datos actuales hacen difícil que estas hipótesis previas puedan ser visualizadas previamente y, aun así, podrían existir otras relaciones en los datos no observadas o no imaginadas. Todo ello requiere una automatización que la minería de datos puede proveer.

Un estudio de minería de datos sobre las compras realizadas en un supermercado de Estados Unidos arrojó la correlación entre la compra de pañales y la compra de cervezas. La historia ilustra una asociación imprevista entre dos productos que el sentido común no habría asociado. Aunque la veracidad de la historia es cuestionable, es un ejemplo ilustrativo que nos permite esclarecer los dominios de la estadística y de la minería de datos. En este caso, los algoritmos computacionales buscaban cualquier tipo de relación entre datos y se encontraron con esta asociación sorprendente. Por el contrario, desde la estadística se habría formulado previamente una hipótesis como: «Hasta qué punto la compra de pañales está correlacionada con la compra de fruta y verdura». Difícilmente una persona habría formulado la hipótesis «Hasta qué punto la compra de pañales está correlacionada con la compra de cerveza», puesto que carece de sentido común.

Pyle nos ofrece otro ejemplo ilustrativo de las diferencias entre la estadística y la minería de datos:

«Una compañía de tarjetas de crédito tenía un equipo de estadísticos cuyo trabajo consistía en descubrir "interacciones interesantes en los datos". Su aproximación era escoger muestras representativas de los datos y buscar interacciones interesantes entre ellos. Algunas de las interacciones fueron estadísticamente significativas y fueron propuestas para acciones de marketing, mientras que otras no fueron significativas y se descartaron.

»La compañía contrató un equipo de expertos en minería de datos para analizar los mismos datos. No empezaron con una muestra de los datos y unas hipótesis a testear. Escogieron un conjunto de datos y usaron algoritmos para extraer las posibles interacciones de interés que podían soportar los datos. Con la lista potencial de interacciones generadas, refinaron el modelo. Los mineros de datos buscaron los factores entre los grupos que fueran responsables de mayor beneficio para la empresa (modelado inferencial) y diseñaron modelos que podían predecir qué tipo de personas podrían ser (modelado predictivo).

»Los estadísticos construyeron modelos de regresión lineales y no lineales, analizaron los residuales y los intervalos de confianza. Los mineros usaron varias técnicas, entre ellas reglas de inducción, árboles de decisión y redes neuronales. Los árboles de decisión fueron los que presentaron mejor rendimiento y se usaron para extraer el conocimiento de los datos. Una de las informaciones que extrajeron es que el 0,1 % de los clientes

presentaban un patrón de alto de beneficio para la empresa. Entre ellos, el 30 % de las personas que compraban equipamiento de esquí se gastaban más de 3.000 dólares en un periodo de 30 días. El resultado se usó para ofrecer campañas de marketing directamente enfocadas a este sector con un retorno de la inversión muy elevado para la empresa.»

Pyle (1999, pág. 488)

Pyle razona que una fluctuación del 0,1 % podría haber sido insignificante para un estadístico, pero fue identificado por algoritmos de minería de datos con un resultado significativo desde el punto de vista comercial. Es así que las aproximaciones de la estadística y la minería de datos presentan enfoques complementarios.

Como ya hemos apuntado anteriormente, querer marcar una línea divisoria entre la estadística y la minería de datos es un enfoque reduccionista. El ejemplo ilustra de forma particular una visión general de la estadística como la verificación de hipótesis y la minería de datos como una versión algorítmica de la búsqueda de estas hipótesis. Si este enfoque nos ayuda a comprender qué es estadística y qué es minería de datos, nos sirve como marco inicial preliminar. Pero no nos gustaría quedarnos limitados por este marco. Veamos qué dicen Witten, Frank y Hall (2011) sobre ello:

«¿Cuál es la diferencia entre aprendizaje automático y estadística? [...] En realidad, no deberíamos buscar una línea divisoria entre aprendizaje automático y estadística, puesto que hay un contínuum –a la vez que multidimensional– de técnicas de análisis de datos. Algunos derivan de las habilidades adquiridas en cursos de estadística y otros están más asociados con los tipos de algoritmos de aprendizaje automático que surgieron con la ciencia computacional. Históricamente, las dos disciplinas han tenido diferentes tradiciones. Si lo forzamos a un único punto diferencial, este sería que la estadística se ha preocupado por testear hipótesis, mientras que el aprendizaje automático se ha preocupado del proceso de generalización formulado como una búsqueda en un espacio de posibles hipótesis. Pero esto es una gran simplificación: la estadística es mucho más que contrastes de hipótesis y muchas técnicas de aprendizaje automático no involucran ningún tipo de búsqueda.

»En el pasado, se han desarrollado varios esquemas a la par en aprendizaje automático y estadística. Uno de ellos son los árboles de inducción. Cuatro estadísticos (Breiman *et al*., 1984) publicaron el libro *Classification and regression trees* a mediados de los años ochenta, y a lo largo de las décadas de los setenta y los ochenta un investigador de aprendizaje automático prominente, J.Ross Quinlan, desarrolló un sistema para inferir árboles de decisión a partir de ejemplos. Los dos proyectos, independientes entre sí, produjeron resultados similares para generar árboles de inducción a partir de ejemplos, y los autores solo fueron conscientes de los trabajos ajenos mucho más tarde.»

Witten, Frank y Hall (2011, pág. 28)

Los autores del libro muestran un buen ejemplo en que la estadística y la minería de datos se dan de la mano. Muchos métodos de minería de datos están diseñados en forma de algoritmos computacionales pero contienen en su interior métodos estadísticos para extraer patrones de conocimiento. Esto se da en todas las fases del proceso de extracción de conocimiento. Las fases de preproceso y preparación de datos contienen técnicas estadísticas como la detección de *outliers* (o valores extremos), la normalización de los datos o la reducción del número de atributos con métodos como el análisis de componentes principales, por citar algunos ejemplos. En la fase de modelado, se usan algoritmos basados en la teoría bayesiana, algoritmos de inducción de árboles usando métricas basadas en entropía o métodos estadísticos para evitar el de-

nominado *overfitting* (sobreaprendizaje). Las técnicas de muestreo estadísticas se usan para obtener estimadores precisos de la calidad de los algoritmos y los contrastes de hipótesis ayudan a identificar los algoritmos que son capaces de extraer los mejores patrones de los datos.

En definitiva, aunque una versión reduccionista de la estadística y la minería de datos establece orígenes distintos y marca diferencias en la formulación y búsqueda de las hipótesis, hay múltiples escenarios en que no solo representan vías de actuación complementarias, sino que además comparten los mismos enfoques.

5. Aplicaciones de la minería de datos

Han y Kamber (2001) también distinguen entre el análisis de datos de tipo descriptivo y de tipo predictivo en el ámbito de la minería de datos. En el análisis descriptivo se caracterizan las propiedades de los datos. Un ejemplo de ello es categorizar los clientes habituales de un supermercado según sus hábitos de compra. En el análisis predictivo se realizan inferencias sobre los datos actuales con el objetivo de realizar predicciones sobre el futuro. Por ejemplo, cómo se comportará el cliente del supermercado si se realiza una campaña de descuentos.

Hay un conjunto finito de tipologías de preguntas que se pueden realizar sobre los datos desde la minería de datos. Entre las más habituales están:

- Caracterización de los datos o discriminación en categorías (clasificación)
- Regresión
- Análisis de asociación
- *Clustering* o agrupación
- Análisis de *outliers*
- Análisis de tendencias o series temporales

Sin intención de adentrarnos en detalle en estas preguntas, a continuación realizamos una breve descripción de este tipo de aplicaciones:

1) Clasificación. La clasificación consiste en la categorización de unos datos en un conjunto de clases predefinido previamente. Habitualmente, se disponen de datos registrados en un fichero o base de datos, caracterizados por un conjunto de atributos y una clase asociada. En términos de estadística, los atributos corresponderían a las variables independientes y la clase asociada correspondería a la variable dependiente. Los algoritmos de clasificación construyen un modelo que explica las relaciones inherentes entre los atributos y su clase. Este proceso se llama también *generalización*, puesto que parte de casos particulares (la muestra) y da como resultado un modelo general que explica los patrones inherentes en dichos datos.

Si tomamos el ejemplo de la tabla 1, los atributos o variables independientes serían los mostrados en la tabla (edad, sexo, estado civil, nivel de ingresos…) y la clase podría ser si el cliente devuelve el crédito en el tiempo y condiciones establecidos. La clasificación construiría un modelo que discriminaría de manera general qué tipo de clientes devuelven el crédito concedido y qué clientes no lo hacen. El modelo podría tener capacidad explicativa y por tanto, usar-

se para comprender qué parámetros y valores influyen en la devolución o no devolución de los créditos concedidos. Además, tiene capacidad predictiva, puesto que puede usarse para predecir el comportamiento de futuros clientes.

2) Regresión. La regresión, también denominada predicción numérica, consiste en construir un modelo que relacione un conjunto de atributos con una variable numérica. Es como la clasificación, donde se relacionan atributos con la clase, siendo la clase un valor numérico en lugar de nominal. Los algoritmos de regresión comparten el mismo objetivo que la regresión estadística, que es el de construir modelos que expliquen las relaciones inherentes entre las variables independientes y las variables dependientes (numéricas). Varían en el tipo de métodos usados y por tanto, en el tipo de modelos construidos; aunque también hay algoritmos de regresión que se basan en los mismos modelos matemáticos que los usados por la estadística.

Volviendo al ejemplo de la tabla 1, hablaríamos de regresión si tomamos como variable independiente el valor del crédito solicitado. Con ello, los algoritmos de regresión tratarían de construir un modelo que explicaría el importe del crédito a partir del resto de características. Evidentemente, la calidad del modelo depende de si existen tales relaciones en los datos y por tanto, sería necesario evaluar el error cometido en la construcción de estos modelos.

3) Análisis de asociación. El análisis de asociaciones consiste en la búsqueda y extracción de modelos que muestran las asociaciones entre atributos de un conjunto de datos. Una forma habitual de representar estas asociaciones es mediante las *reglas de asociación*. En este tipo de enfoque, no hay una clase o variable dependiente identificada *a priori*, como sucede en la clasificación y regresión. Este enfoque trata todas las variables «por igual» e investiga si existe cualquier tipo de relación entre ellas.

De nuevo, usando el ejemplo de la tabla 1, las reglas de asociación podrían ser del tipo: si el número de hijos es 2 o 3 y el estado civil es casado está relacionado con un crédito para vehículo o reformas y el importe oscila entre 20.000 y 40.000. Como se ve, el modelo contiene valor descriptivo en forma de reglas. Al igual que en los casos anteriores, la calidad de estos modelos depende tanto de la existencia de estas relaciones inherentes en los datos como de la capacidad del modelo de identificar este tipo de relaciones.

Como se comentaba anteriormente, el enfoque de la estadística para hallar este tipo de asociaciones parte de hipótesis previamente establecidas. Por ejemplo, se podría formular si el número de hijos está relacionado con el importe del crédito. Pero estas asociaciones deben ser imaginadas a priori. Mediante la minería de datos, el algoritmo busca «a ciegas» qué tipo de relaciones se pueden extraer. El algoritmo realiza una búsqueda de las hipótesis más promete-

doras dentro del espacio de todas las posibles hipótesis. Puede imaginarse el estudiante que tal búsqueda a ciegas necesita una gran capacidad de cómputo para poderse llevar a cabo.

4) Agrupación (*clustering*). La agrupación consiste en separar los datos en distintos grupos o categorías. Cabe distinguirla de la clasificación, donde las categorías de la muestra están asignadas *a priori*. En la agrupación no hay categorías preestablecidas y es el propio algoritmo el que agrupa los datos según sus características, de manera que los casos de una misma categoría presenten un grado de similitud elevado y a la vez se diferencien del resto de casos de los otros grupos.

La agrupación se denomina también segmentación o en inglés, *clustering*. En el caso de la tabla 1, la agrupación revelaría qué tipos de clientes forman parte del conjunto de datos. Por ejemplo, se podría extraer que un segmento de datos está formado por personas solteras, de entre 20 y 30 años, que solicitan un préstamo para alquilar una vivienda. Otro grupo de datos podría ser el de personas con familia, casados, con una hipoteca y que piden un préstamo entre 20.000 y 40.000 euros para comprar un vehículo o hacer reformas en casa.

5) Análisis de *outliers*. En los enfoques descritos previamente, los algoritmos de clasificación, regresión o agrupación tratan de encontrar las regularidades en los datos, es decir, modelos que son capaces de describir el comportamiento general de los datos. En el análisis de *outliers* (también denominados valores extremos) se trata precisamente de lo contrario: encontrar los casos anómalos, las rarezas, dentro de los patrones generales. Tanto los algoritmos descritos como la estadística suelen eliminar los datos anómalos porque pueden introducir desviaciones en los análisis que alteran las conclusiones. En cambio, el análisis de valores extremos pretende precisamente descubrir estos casos anómalos porque son el objeto de interés del estudio. Hablamos por ejemplo de la detección de fraude en transacciones de tarjetas de crédito o la detección de petróleo vertido en océanos.

6) Análisis de series temporales. En el análisis de series temporales, los datos bajo estudio tienen un componente temporal. Nos referimos por ejemplo a la evolución de la cotización del IBEX 35 en los últimos diez años, o la temperatura máxima y mínima diaria de Madrid a lo largo de los tres últimos años. El objetivo del estudio de las series temporales es modelar el comportamiento de estos datos a lo largo del tiempo para predecir el comportamiento futuro. Como podrá intuir el estudiante, la predicción del comportamiento futuro no es tarea nada fácil, puesto que no es posible identificar todos los datos que influyen en este comportamiento (imaginemos todas las variables que pueden influir en la cotización del IBEX 35). En definitiva, el modelado de fenómenos complejos de la realidad no es tarea sencilla.

Resumen

En este módulo hemos navegado por la superficie del análisis de datos, y en especial de la estadística. Decimos «navegar» puesto que hemos realizado un recorrido, aunque no exhaustivo, por sus orígenes y nos hemos detenido en algunas de sus anécdotas, con el ánimo de despertar el interés y la curiosidad del estudiante. El recorrido ha visitado las principales aplicaciones de la estadística sobre la extracción de conclusiones de los datos, el manejo de la incertidumbre, el análisis de las relaciones entre variables, el muestreo, la predicción y la toma de decisiones. Asimismo, la distinción entre estadística descriptiva e inferencial es clave para comprender si se extraen conclusiones sobre la muestra de estudio o inferencias sobre la población.

El análisis de datos se configura hoy en día como un panorama multidisciplinar. Por este motivo es conveniente contextualizar la estadística en este marco y, a la vez, contrastarla con la joven disciplina de la minería de datos. Hemos descrito brevemente los principales enfoques de la minería de datos, como la clasificación, regresión, asociación, agrupación, análisis de *outliers* y series temporales. Estadística y minería de datos provienen de enfoques distintos, de carácter matemático-científico la primera y computacional la segunda. Aun así comparten el objetivo de extraer información útil de los datos que ayuden a comprender los fenómenos que nos rodean en un amplio espectro de dominios, como la ciencia, la medicina, la meteorología, la economía y la empresa, por citar tan solo algunos de ellos.

Después de este breve recorrido animamos al estudiante a profundizar en las técnicas y métodos de la estadística, puesto que con ello encontrará los fundamentos del análisis de datos.

Bibliografía

Alea, V.; Guillén, M.; Muñoz, M. C.; Torrelles, E.; Viladomiu, N. (1999). *Estadística aplicada a les ciències econòmiques i socials*. Barcelona: Edicions Universitat de Barcelona, Mc-Graw-Hill.

Armitage, P.; Berry, G.; Matthews, J. N. S. (2002). *Statistical Methods in Medical Research (Fourth Edition)*. Malden, Massachusets: Blackwell Science.

Dodge, Y. (2006). *The Oxford Dictionary of Statistical Terms*. OUP.

Drucker, P. F. (2006). *Innovation and Entrepreneurship. Practice and Principles*. HarperCollins.

Gibergans, J.; Gil, A. J.; Rovira, C. (2009). *Estadística*. Barcelona: FUOC.

Han, J.; Kamber, M. (2001). *Data Mining.Concepts and Techniques*. San Diego: Academic Press.

Mayer-Schönberger, V.; Cukier, K. (2013). *Big Data. A Revolution That Will Transform How We Live, Work and Think*. UK: John Murray.

Moore, D. S.; McCabe, G. P.; Craig, B. A. (2012). *Introduction to the Practice of Statistics* (7ª edición). New York: W.H. Freeman and Company.

Newbold, P. (1997). *Estadística para los negocios y la Economía* (4ª edición). Pearson, Prentice Hall.

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann Publishers, Inc.

Rooney, A. (2009). *Historia de las matemáticas. De la construcción de las pirámides hasta la exploración del infinito*. Barcelona: Oniro.

Rothman, K. J. (1996). *Lessons from John Graunt. The Lancet*. Elsevier.

Scheaffer, R. L. (1999). *Sampling Methods and Practice. NCSSM Statistics Leadership Institute*. University of Florida.

Thomas, D. B. (1991). «The WHO collaborative study of neoplasia and steroid contraceptives: The influence of combined oral contraceptives on risk of neoplasms in developing and developed countries». *Contraception* (43 (6), págs. 695-710). Disponible en:

<https://doi.org/10.1016/0010-7824(91)90010-D>

Tomeo, V.; Uña, I. (2003). *Lecciones de Estadística descriptiva. Curso teórico-práctico*. Madrid: Thomson.

Vigen, T.*(2015).Spurious correlations*. Hachette Books.

Witten, I.; Frank, E.; Hall, M. A. (2011). *Data Mining. Practical Machine Learning Tools and Techniques*.(3ª edición). Burlington: Morgan Kaufmann.