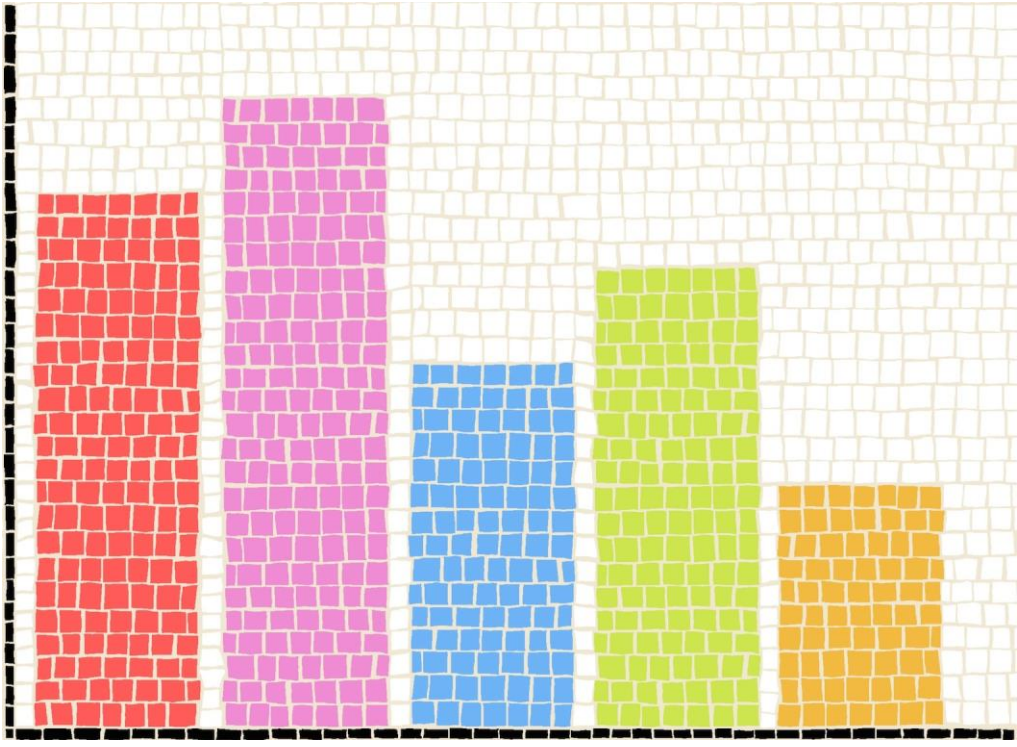
El arte de la ciencia de los datos

Una guía para todos los que trabajan con datos



Roger D. Peng y Elizabeth Matsu i

El arte de la ciencia de los datos

Una guía para cualquiera que trabaje con datos

Roger D. Peng y Elizabeth Matsui

Este libro está a la venta en
http://leanpub.com/artofdatascience

Esta versión fue publicada el 2015-09-28

También por Roger D. Peng

Programación en R para la

ciencia de los datos Análisis

exploratorio de datos con R

Redacción de informes para la ciencia de datos en R

Un agradecimiento especial a Maggie Matsui, que creó todas las ilustraciones de este libro.

Contenido

1. El análisis de datos como arte

El análisis de datos es difícil, y parte del problema es que pocas personas pueden explicar cómo hacerlo. No es que no haya gente que haga análisis de datos con regularidad. Es que las personas que son realmente buenas en ello todavía tienen que ilustrarnos sobre el proceso de pensamiento que pasa por sus cabezas.

Imagina que le preguntas a una compositora cómo escribe sus canciones. Hay muchas herramientas a las que puede recurrir. Tenemos una idea general de cómo debe estructurarse una buena canción: qué longitud debe tener, cuántas estrofas, si hay una estrofa seguida de un estribillo, etc. En otras palabras, hay un marco abstracto para las canciones en general. Del mismo modo, tenemos la teoría musical que nos dice que ciertas combinaciones de notas y acordes funcionan bien juntos y otras combinaciones no suenan bien. Por muy buenas que sean estas herramientas, en última instancia, el conocimiento de la estructura de la canción y la teoría musical por sí solas no permiten crear una buena canción. Se necesita algo más.

En el legendario ensayo de Donald Knuth de 1974 *La programación informática como arte*[1]Knuth habla de la diferencia entre arte y ciencia. En ese ensayo, intentaba transmitir la idea de que, aunque la programación de ordenadores implicaba máquinas complejas y conocimientos muy técnicos, el acto de escribir un programa informático tenía un componente artístico. En este ensayo, dice que

> La ciencia es un conocimiento que entendemos tan bien que podemos enseñárselo a un ordenador.

[1]http://www.paulgraham.com/knuth.html

Todo lo demás es arte.

En algún momento, el compositor debe inyectar una chispa creativa en el proceso para reunir todas las herramientas de composición y hacer algo que la gente quiera escuchar. Esta es una parte clave del *arte* de la composición. Esa chispa creativa es difícil de describir, y mucho más de escribir, pero es claramente esencial para escribir buenas canciones. Si no fuera así, tendríamos programas informáticos que escribirían regularmente canciones de éxito. Para bien o para mal, eso aún no ha ocurrido.

Al igual que la composición de canciones (y la programación de ordenadores, por cierto), es importante darse cuenta de que *el análisis de datos es un arte*. Todavía no es algo que podamos enseñar a un ordenador. Los analistas de datos tienen muchas *herramientas* a su disposición, desde la regresión lineal hasta los árboles de clasificación e incluso el aprendizaje profundo, y todas estas herramientas han sido cuidadosamente enseñadas a los ordenadores. Pero, en última instancia, un analista de datos debe encontrar la manera de reunir todas las herramientas y aplicarlas a los datos para responder a una pregunta relevante, una pregunta de interés para las personas.

Por desgracia, el proceso de análisis de datos no es algo que hayamos podido escribir de forma eficaz. Es cierto que existen muchos libros de texto de estadística, muchos de los cuales llenan nuestras propias estanterías. Pero, en nuestra opinión, ninguno de ellos aborda realmente los problemas centrales que implica la realización de análisis de datos en el mundo real. En 1991, Daryl Pregibon, un pionero estadístico que trabajó en AT&T Research y ahora en Google, dijo en referencia al proceso de análisis de datos[2] que "los estadísticos tienen un proceso que defienden pero que no entienden del todo".

La descripción del análisis de datos presenta un difícil dilema. Por un lado, el desarrollo de un marco útil implica la caracterización de los elementos de un análisis de datos

cambiarlo
mediante el uso de un sistema abstracto.

[2] http://www.nap.edu/catalog/1910/the-future-of-statistical-software- actas de un foro

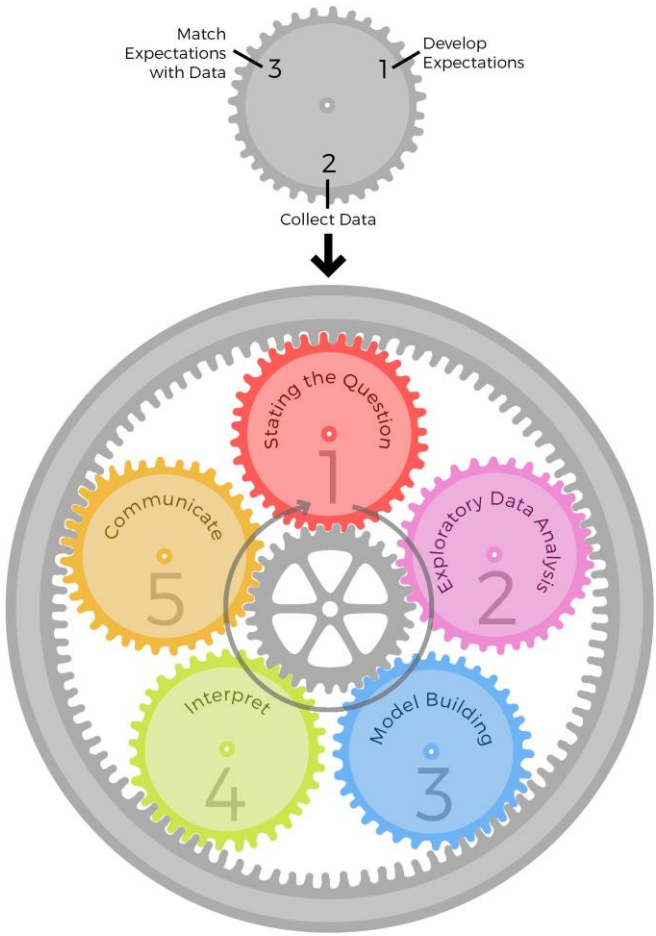
lenguaje con el fin de encontrar los puntos comunes entre los diferentes tipos de análisis. A veces, este lenguaje es el de las matemáticas. Por otra parte, a menudo son los propios detalles de un análisis los que hacen que cada uno sea tan difícil y, a la vez, interesante. ¿Cómo se puede generalizar eficazmente entre muchos análisis de datos diferentes, cada uno de los cuales tiene aspectos únicos importantes?

Lo que nos hemos propuesto hacer en este libro es escribir el proceso de análisis de datos. Lo que describimos no es una "fórmula" específica para el análisis de datos -algo así como "aplique este método y luego haga esa prueba"- sino que es un proceso general que puede aplicarse en una variedad de situaciones. A través de nuestra amplia experiencia tanto en la gestión de analistas de datos como en la realización de nuestros propios análisis de datos, hemos observado detenidamente lo que produce resultados coherentes y lo que no produce una visión útil de los datos. Nuestro objetivo es escribir lo que hemos aprendido con la esperanza de que otros puedan encontrarlo útil.

2. Epiciclos de análisis

Para los no iniciados, un análisis de datos puede parecer que sigue un proceso lineal, de un paso tras otro, que al final llega a un resultado bien empaquetado y coherente. En realidad, el análisis de datos es un proceso altamente iterativo y no lineal, que se refleja mejor en una serie de epiciclos (véase la figura), en los que se aprende información en cada paso, que luego informa sobre si (y cómo) perfeccionar, y rehacer, el paso que se acaba de realizar, o si (y cómo) proceder al siguiente paso.

Un epiciclo es un pequeño círculo cuyo centro se mueve alrededor de la circunferencia de un círculo mayor. En el análisis de datos, el proceso iterativo que se aplica a todos los pasos del análisis de datos puede concebirse como un epiciclo que se repite para cada paso a lo largo de la circunferencia de todo el proceso de análisis de datos. Algunos análisis de datos parecen ser fijos y lineales, como los algoritmos integrados en diversas plataformas de software, incluidas las aplicaciones. Sin embargo, estos algoritmos son productos finales de análisis de datos que han surgido del trabajo no lineal de desarrollar y refinar un análisis de datos para que pueda ser "algoritmizado".



Epiciclos de análisis

2.1 Preparando el escenario

Antes de adentrarnos en el "epiciclo de análisis", conviene detenerse a considerar qué entendemos por "análisis de datos". Aunque muchos de los conceptos que discutiremos en este

El marco y los conceptos de este capítulo y de los siguientes son aplicables a la realización de un *estudio*, pero se adaptan específicamente a la realización de un *análisis de datos*. Mientras que un estudio incluye el desarrollo y la ejecución de un plan de recogida de datos, un análisis de datos presupone que los datos ya se han recogido. Más concretamente, un estudio incluye el desarrollo de una hipótesis o pregunta, el diseño del proceso de recogida de datos (o protocolo del estudio), la recogida de datos y el análisis e interpretación de los mismos. Dado que un análisis de datos presupone que los datos ya se han recogido, incluye el desarrollo y el perfeccionamiento de una pregunta y el proceso de análisis e interpretación de los datos. Es importante señalar que, aunque el análisis de datos suele realizarse sin llevar a cabo un estudio, también puede llevarse a cabo como componente de un estudio.

2.2 Epiciclo de análisis

Hay 5 actividades principales de análisis de datos:

1. Plantear y afinar la pregunta
2. Exploración de los datos
3. Construcción de modelos estadísticos formales
4. Interpretación de los resultados
5. Comunicar los resultados

Estas 5 actividades pueden tener lugar en diferentes escalas de tiempo: por ejemplo, puede pasar por las 5 en el transcurso de un día, pero también ocuparse de cada una, en el caso de un gran proyecto, a lo largo de muchos meses. Antes de hablar de estas actividades principales, que se desarrollarán en capítulos posteriores, será importante entender primero el marco general utilizado para abordar cada una de estas actividades.

Aunque hay muchos tipos diferentes de actividades que se pueden llevar a cabo durante el análisis de datos, todos los aspectos del proceso completo se pueden abordar a través de un proceso interactivo que llamamos "epiciclo de análisis de datos". Más concretamente, para cada una de las cinco actividades principales, es fundamental que realice los siguientes pasos:

1. Establecer expectativas,
2. Recoger información (datos), comparar los datos con sus expectativas, y si las expectativas no coinciden,
3. Revisar sus expectativas o arreglar los datos para que sus datos y sus expectativas coincidan.

La iteración a través de este proceso de 3 pasos es lo que llamamos el "epiciclo del análisis de datos". Al pasar por cada etapa de un análisis, tendrá que recorrer el epiciclo para perfeccionar continuamente su pregunta, su análisis exploratorio de datos, sus modelos formales, su interpretación y su comunicación.

El ciclo repetido a través de cada una de estas cinco actividades básicas que se realiza para completar un análisis de datos forma el círculo mayor del análisis de datos (véase la figura). En este capítulo explicamos en detalle en qué consiste este proceso epicicloidal de 3 pasos y damos ejemplos de cómo puede aplicarlo a su análisis de datos.

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Epiciclos de análisis

2.3 Fijar las expectativas

El desarrollo de expectativas es el proceso de pensar deliberadamente en lo que se espera antes de hacer algo, como inspeccionar los datos, realizar un procedimiento o introducir un mandato. Para los analistas de datos experimentados, en algunas circunstancias, el desarrollo de las expectativas puede ser un proceso automático, casi subconsciente, pero es una actividad importante que hay que cultivar y ser deliberada.

Por ejemplo, puede que vaya a cenar con unos amigos a un establecimiento que sólo dispone de efectivo y tenga que pasar por el cajero automático para sacar dinero antes de reunirse. Para tomar una decisión sobre la cantidad de dinero que va a retirar, tiene que haber desarrollado alguna expectativa sobre el coste de la cena. Puede ser una expectativa automática porque cenas en ese establecimiento con regularidad, así que sabes cuál es el

El coste típico de una comida está ahí, lo que sería un ejemplo de conocimiento *a priori*. Otro ejemplo de conocimiento *a priori* sería saber lo que cuesta una comida típica en un restaurante de tu ciudad, o saber lo que cuesta una comida en los restaurantes más caros de tu ciudad. A partir de esta información, se podría establecer un límite superior e inferior del coste de la comida.

Es posible que también haya buscado información externa para desarrollar sus expectativas, lo que podría incluir preguntar a sus amigos que se reunirán con usted o que han comido en el restaurante antes y/o buscar en Google el restaurante para encontrar información general sobre el coste en línea o un menú con precios. Este mismo proceso, en el que utilizas cualquier información a priori que tengas y/o fuentes externas para determinar lo que esperas cuando inspeccionas tus datos o ejecutas un procedimiento de análisis, se aplica a cada actividad central del proceso de análisis de datos.

2.4 Recogida de información

Este paso implica la recopilación de información sobre su pregunta o sus datos. Para la pregunta, se recopila información realizando una búsqueda bibliográfica o preguntando a expertos para asegurarse de que la pregunta es buena. En el próximo capítulo hablaremos de las características de una buena pregunta. En cuanto a los datos, después de tener algunas expectativas sobre el resultado que se obtendrá al inspeccionar los datos o realizar el procedimiento de análisis, se realiza la operación. Los resultados de esa operación son los datos que necesita recoger, y luego determina si los datos recogidos coinciden con sus expectativas. Para ampliar la metáfora del restaurante, cuando se va al restaurante, pedir la cuenta es recoger los datos.

2.5 Comparación de las expectativas con los datos

Ahora que tiene los datos en la mano (la cuenta del restaurante), el siguiente paso es comparar sus expectativas con los datos. Hay dos resultados posibles: o bien tus expectativas del coste coinciden con el importe de la cuenta, o bien no coinciden. Si tus expectativas y los datos coinciden, estupendo, puedes pasar a la siguiente actividad. Si, por el contrario, tus expectativas eran un coste de 30 dólares, pero el cheque era de 40 dólares, tus expectativas y los datos no coinciden. Hay dos posibles explicaciones para la discordancia: en primer lugar, tus expectativas eran erróneas y hay que revisarlas, o en segundo lugar, el cheque era erróneo y contiene un error. Revisas la cuenta y descubres que te han cobrado dos postres en lugar de uno, y concluyes que hay un error en los datos, por lo que pides que se corrija la cuenta.

Un indicador clave de lo bien que va el análisis de los datos es lo fácil o difícil que resulta hacer coincidir los datos recogidos con las expectativas iniciales. Hay que configurar las expectativas y los datos de manera que resulte fácil hacerlos coincidir. En el ejemplo del restaurante, su expectativa era de 30 dólares y los datos indicaban que la comida costaba 40 dólares, por lo que es fácil ver que (a) su expectativa estaba fuera de lugar por 10 dólares y que (b) la comida era más cara de lo que pensaba. Cuando vuelvas a este lugar, puede que lleves 10 dólares más. Si nuestra expectativa original era que la comida costaría entre 0 y
1.000 dólares, entonces es cierto que nuestros datos entran en ese rango, pero no está claro cuánto más hemos aprendido. Por ejemplo, ¿cambiaría su comportamiento la próxima vez que volviera? La expectativa de una comida de 30 dólares se denomina a veces hipótesis tajante porque afirma algo muy concreto que puede verificarse con los datos.

2.6 Aplicación del proceso de epígrafe de análisis

Antes de hablar de un par de ejemplos, vamos a repasar los tres pasos que hay que seguir para cada actividad de análisis de datos básicos. Estos son:

1. Fijar las expectativas,
2. Recoger información (datos), comparar los datos con sus expectativas, y si las expectativas no coinciden,
3. Revisar sus expectativas o arreglar los datos para que sus expectativas y los datos coincidan.

Ejemplo: Prevalencia del asma en EE.UU.

Apliquemos el "epiciclo de análisis de datos" a un ejemplo muy básico. Supongamos que su pregunta inicial es determinar la prevalencia del asma entre los adultos, porque su empresa quiere saber cuál puede ser el tamaño del mercado para un nuevo medicamento contra el asma. Tiene una pregunta general que ha sido identificada por su jefe, pero necesita (1) afinar la pregunta, (2) explorar los datos, (3) construir un modelo estadístico, (4) interpretar los resultados y (5) comunicar los resultados. Aplicaremos el "epiciclo" a cada una de estas cinco actividades fundamentales.

En la primera actividad, el perfeccionamiento de la pregunta, primero se desarrollan las expectativas de la pregunta, luego se recopila información sobre la pregunta y se determina si la información recopilada coincide con las expectativas, y si no, se revisa la pregunta. Sus expectativas son que la respuesta a esta pregunta es desconocida y que la pregunta es contestable. Sin embargo, una búsqueda bibliográfica y en Internet revela que esta pregunta ha sido respondida (y que los Centros de Control de Enfermedades (CDC) la responden continuamente), por lo que usted

reconsiderar la pregunta, ya que basta con ir al sitio web de los CDC para obtener datos recientes sobre la prevalencia del asma.

Informas a tu jefe e inicias una conversación en la que se pone de manifiesto que cualquier nuevo fármaco que se desarrolle tar- bará a aquellos cuyo asma no se controla con la medicación actualmente disponible, por lo que identificas una pregunta mejor, que es "¿cuántas personas en Estados Unidos tienen asma que no está actualmente controlada, y cuáles son los predictores demográficos del asma no controlada?" Repite el proceso de recopilación de información para determinar si su pregunta puede responderse y es una buena pregunta, y continúa este proceso hasta que esté satisfecho de haber refinado su pregunta de modo que tenga una buena pregunta que pueda responderse con los datos disponibles.

Supongamos que ha identificado una fuente de datos que puede descargarse de un sitio web y que es una muestra que representa a la población adulta de Estados Unidos, mayor de 18 años. La siguiente actividad es el análisis exploratorio de datos, y usted comienza con la expectativa de que cuando inspeccione sus datos habrá 10.123 filas (o registros), cada una de las cuales representa a un individuo en los Estados Unidos, ya que esta es la información proporcionada en la documentación, o libro de códigos, que viene con el conjunto de datos. El libro de códigos también le indica que habrá una variable que indique la edad de cada individuo en el conjunto de datos.

Sin embargo, cuando inspecciona los datos, se da cuenta de que sólo hay 4.803 filas, así que vuelve al libro de códigos para confirmar que sus expectativas son correctas sobre el número de filas, y cuando confirma que sus expectativas son correctas, vuelve al sitio web donde descargó los archivos y descubre que había dos archivos que contenían los datos que necesitaba, con un archivo que contenía 4.803 registros y el segundo archivo que contenía los 5.320 registros restantes. Descarga el segundo archivo y lo lee en su programa estadístico

paquete de software y añadir el segundo archivo al primero.

Ahora tiene el número correcto de filas, así que pasa a determinar si sus expectativas sobre la edad de la población coinciden con sus expectativas, que es que todos tienen 18 años o más. Resume la variable edad, de modo que puede ver los valores mínimos y máximos y encuentra que todos los individuos tienen 18 años o más, lo que coincide con sus expectativas. Aunque hay más cosas que podrías hacer para inspeccionar y explorar tus datos, estas dos tareas son ejemplos del enfoque a seguir. En última instancia, usted utilizará este conjunto de datos para estimar la prevalencia del asma no controlada entre los adultos en los Estados Unidos.

La tercera actividad es la construcción de un modelo estadístico, necesario para determinar las características demográficas que mejor predicen que alguien tiene asma no controlada. Los modelos estadísticos sirven para producir una formulación precisa de su pregunta, de modo que pueda ver exactamente cómo quiere utilizar sus datos, ya sea para estimar un parámetro específico o para hacer una predicción. Los modelos estadísticos también proporcionan un marco formal en el que puede cuestionar sus hallazgos y poner a prueba sus suposiciones.

Ahora que ha calculado la prevalencia del asma no controlada entre los adultos de EE.UU. y ha determinado que la edad, el sexo, la raza, el índice de masa corporal, el hábito de fumar y los ingresos son los mejores predictores disponibles del asma no controlada, pasará a la cuarta actividad principal, que es la interpretación de los resultados. En realidad, la interpretación de los resultados se realiza al mismo tiempo que se construye el modelo y también después de haber terminado de construirlo, pero conceptualmente son actividades distintas.

Supongamos que ha construido su modelo final y que pasa a interpretar los resultados de su modelo. Cuando examina su modelo predictivo final, inicialmente sus expectaciones se corresponden con la edad, la raza afroamericana/negra,

El índice de masa corporal, el hábito de fumar y los bajos ingresos se asocian positivamente con el asma no controlada.

Sin embargo, se da cuenta de que el sexo femenino está *inversamente* asociado al asma no controlada, cuando su investigación y las conversaciones con los expertos indican que, entre los adultos, el sexo femenino debería estar positivamente asociado al asma no controlada. Este desajuste entre las expectativas y los resultados le lleva a detenerse y explorar un poco para determinar si sus resultados son realmente correctos y debe ajustar sus expectativas o si hay un problema con sus resultados más que con sus expectativas. Después de indagar un poco, descubres que habías pensado que la variable de género se codificaba como 1 para las mujeres y 0 para los hombres, pero en cambio el libro de códigos indica que la variable de género se codificaba como 1 para los hombres y 0 para las mujeres. Por tanto, la interpretación de sus resultados era incorrecta, no sus expectativas. Ahora que entiende cuál es la codificación de la variable de género, su interpretación de los resultados del modelo coincide con sus expectativas, por lo que puede pasar a comunicar sus resultados.

Por último, comunicas tus conclusiones, y sí, la epicidad se aplica también a la comunicación. A efectos de este ejemplo, supongamos que has elaborado un informe informal que incluye un breve resumen de tus conclusiones. Su expectativa es que su informe comunique la información que a su jefe le interesa conocer. Te reúnes con tu jefa para revisar los resultados y ella te hace dos preguntas:
(1) la fecha de recopilación de los datos en el conjunto de datos y (2) cómo se espera que los patrones demográficos cambiantes que se prevén en los próximos 5-10 años afecten a la prevalencia del asma no controlada. Aunque puede ser decepcionante que su informe no satisfaga plenamente las necesidades de su jefe, recibir comentarios es una parte fundamental de la realización de un análisis de datos y, de hecho, diríamos que un buen análisis de

datos requiere comunicación, comentarios y, a continuación

acciones en respuesta a los comentarios.

Aunque conoces la respuesta sobre los años en que se recogieron los datos, te das cuenta de que no incluiste esta información en tu informe, así que lo revisas para incluirla. También te das cuenta de que la pregunta de tu jefe sobre el efecto de los cambios demográficos en la prevalencia del asma no controlada es buena, ya que tu empresa quiere predecir el tamaño del mercado en el futuro, así que ahora tienes un nuevo análisis de datos que abordar. También debería sentirse bien por el hecho de que su análisis de datos haya sacado a la luz nuevas preguntas, ya que ésta es una de las características de un análisis de datos exitoso.

En los próximos capítulos, utilizaremos ampliamente este marco para discutir cómo cada actividad del proceso de análisis de datos debe repetirse continuamente. Aunque la ejecución de los tres pasos puede parecer tediosa al principio, con el tiempo se le cogerá el tranquillo y el ciclo del proceso se producirá de forma natural e inconsciente. De hecho, nos atreveríamos a decir que la mayoría de los mejores analistas de datos ni siquiera se dan cuenta de que lo están haciendo.

3. Plantear y afinar la pregunta

El análisis de datos requiere un poco de reflexión y creemos que, cuando se ha realizado un buen análisis de datos, se ha pasado más tiempo pensando que haciendo. La reflexión comienza incluso antes de mirar un conjunto de datos, y merece la pena dedicar una cuidadosa reflexión a la pregunta. Nunca se insistirá lo suficiente en este punto, ya que muchos de los escollos "fatales" de un análisis de datos pueden evitarse si se dedica la energía mental necesaria para formular correctamente la pregunta. En este capítulo, hablaremos de las características de una buena pregunta, de los tipos de preguntas que pueden formularse y de cómo aplicar el proceso epicicloidal iterativo para formular y perfeccionar su pregunta, de modo que cuando empiece a examinar los datos, tenga una pregunta nítida y con respuesta.

3.1 Tipos de preguntas

Antes de entrar en la formulación de la pregunta, es útil considerar cuáles son los diferentes tipos de preguntas. Hay seis tipos básicos de preguntas y gran parte del análisis que sigue procede de un artículo[1] publicado en *Science* por Roger y Jeff Leek[2]. Comprender el tipo de pregunta que se formula puede ser el paso más fundamental para garantizar que, al final, la interpretación de los resultados sea correcta. Los seis tipos de preguntas son:

[1] http://www.sciencemag.org/content/347/6228/1314.short
[2] http://jtleek.com

1. Descriptivo
2. Exploración
3. Inferencial
4. Predictivo
5. Causal
6. Mecánica

Además, el tipo de pregunta que se formula influye directamente en la interpretación de los resultados.

Una pregunta *descriptiva* es aquella que pretende resumir una característica de un conjunto de datos. Los ejemplos incluyen la determinación de la proporción de hombres, el número medio de porciones de frutas y verduras frescas por día, o la frecuencia de enfermedades virales en un conjunto de datos recogidos de un grupo de individuos. No hay interpretación del resultado en sí mismo, ya que el resultado es un hecho, un atributo del conjunto de datos con el que se está trabajando.

Una pregunta *exploratoria* es aquella en la que se analizan los datos para ver si existen patrones, tendencias o relaciones entre las variables. Estos tipos de análisis también se denominan análisis "generadores de hipótesis" porque, en lugar de probar una hipótesis como se haría con una pregunta inferencial, causal o mecanicista, se buscan patrones que apoyen la propuesta de una hipótesis. Si se tiene la idea general de que la dieta está relacionada de alguna manera con las enfermedades víricas, se podría explorar esta idea examinando las relaciones entre una serie de factores dietéticos y las enfermedades víricas. En su análisis exploratorio encuentra que los individuos que consumen una dieta rica en ciertos alimentos tienen menos enfermedades víricas que aquellos cuya dieta no está enriquecida con estos alimentos, por lo que propone la hipótesis de que, entre los adultos, el consumo de al menos 5 raciones al día de fruta y verdura fresca está asociado con menos enfermedades víricas al año.

Una pregunta *inferencial* sería un replanteamiento de esta hipótesis planteada como una pregunta y se respondería analizando un conjunto diferente de datos, que en este ejemplo, es una muestra representativa de adultos en EE.UU. Al analizar este conjunto diferente de datos, usted está determinando si la asociación que observó en su análisis exploratorio se mantiene en una muestra diferente y si se mantiene en una muestra que es representativa de la población adulta de EE.UU., lo que sugeriría que la asociación es aplicable a todos los adultos de EE.UU.. En otras palabras, podrá inferir lo que es cierto, en promedio, para la población adulta de EE.UU. a partir del análisis que realice en la muestra representativa.

Una pregunta *predictiva* sería aquella en la que se pregunta qué tipo de personas consumirán una dieta rica en frutas y verduras frescas durante el próximo año. En este tipo de pregunta, no le interesa tanto lo que hace que alguien siga una determinada dieta, sino lo que predice si alguien seguirá esa dieta. Por ejemplo, los ingresos más altos pueden ser uno de los predictores finales, y es posible que no sepa (ni le importe) por qué las personas con ingresos más altos son más propensas a seguir una dieta rica en frutas y verduras frescas, pero lo más importante es que los ingresos son un factor que predice este comportamiento.

Aunque una pregunta inferencial podría decirnos que las personas que comen un determinado tipo de alimentos tienden a tener menos enfermedades víricas, la respuesta a esta pregunta no nos dice si el consumo de estos alimentos provoca una reducción del número de enfermedades víricas, lo que sería el caso de una pregunta *causal*. Una pregunta causal se refiere a si el cambio de un factor cambiará otro factor, en promedio, en una población. A veces, el diseño subyacente de la recogida de datos permite, por defecto, que la pregunta que se formula sea causal. Un ejemplo de ello serían los datos recogidos en el contexto de un ensayo aleatorio, en el que las personas fueron asignadas al azar a seguir una dieta rica en frutas y verduras frescas o una que

era bajo en frutas y verduras frescas. En otros casos, aunque los datos no procedan de un ensayo aleatorio, se puede adoptar un enfoque analítico diseñado para responder a una pregunta causal.

Por último, ninguna de las preguntas descritas hasta ahora nos llevará a una respuesta que nos diga, si la dieta provoca, efectivamente, una reducción del número de enfermedades víricas, *cómo* la dieta conduce a una reducción del número de enfermedades víricas. Una pregunta que se refiera a cómo una dieta rica en frutas y verduras frescas conduce a una reducción del número de enfermedades víricas sería una pregunta *mecanicista*.

Hay un par de puntos adicionales sobre los tipos de preguntas que son importantes. En primer lugar, por necesidad, muchos análisis de datos responden a múltiples tipos de preguntas. Por ejemplo, si un análisis de datos pretende responder a una pregunta inferencial, las preguntas descriptivas y exploratorias también deben responderse durante el proceso de respuesta a la pregunta inferencial. Siguiendo con nuestro ejemplo de la dieta y las enfermedades víricas, no se pasaría directamente a un modelo estadístico de la relación entre una dieta rica en frutas y verduras frescas y el número de enfermedades víricas sin haber determinado la frecuencia de este tipo de dieta y de las enfermedades víricas y su relación entre sí en esta muestra. Un segundo punto es que el tipo de pregunta que se formula está determinado en parte por los datos de que se dispone (a menos que se planee realizar un estudio y recoger los datos necesarios para hacer el análisis). Por ejemplo, es posible que quiera formular una pregunta causal sobre la dieta y las enfermedades víricas para saber si una dieta con alto contenido en frutas y verduras frescas provoca una disminución del número de enfermedades víricas, y el mejor tipo de datos para responder a esta pregunta causal es aquel en el que la dieta de las personas cambia de una con alto contenido en frutas y verduras frescas a otra que no lo es, o viceversa. Si no existe este tipo de datos, entonces lo mejor que se puede hacer es aplicar la causalidad

métodos de análisis a los datos observacionales o, por el contrario, responder a una pregunta inferencial sobre la dieta y las enfermedades virales.

3.2 Aplicar el epiciclo a la formulación y el perfeccionamiento de la pregunta

Ahora puede utilizar la información sobre los tipos de preguntas y las características de las buenas preguntas como guía para refinar su pregunta. Para ello, puedes repetir los tres pasos de:

1. Establecer sus expectativas sobre la pregunta
2. Recopilación de información sobre su pregunta
3. Determinar si las expectativas coinciden con la información recopilada y, a continuación, refinar la pregunta (o las expectativas) si las expectativas no coinciden con la información recopilada.

3.3 Características de una buena pregunta

Hay cinco características clave de una buena pregunta para un análisis de datos, que van desde la característica muy básica de que la pregunta no debe haber sido ya respondida hasta la más abstracta de que cada una de las posibles respuestas a la pregunta debe tener una única interpretación y ser significativa. A continuación, analizaremos con más detalle cómo evaluar esto.

Para empezar, la pregunta debe ser de **interés** para su audiencia, cuya identidad dependerá del contexto y el entorno en el que trabaje con los datos. Si trabaja en el ámbito académico, el público puede ser sus colaboradores, la comunidad científica, los reguladores gubernamentales, su

financiadores y/o el público. Si trabaja en una empresa de nueva creación, su público es su jefe, la dirección de la empresa y los inversores. Por ejemplo, responder a la pregunta de si la contaminación por partículas en el exterior está asociada a problemas de desarrollo en los niños puede interesar a las personas que se dedican a regular la contaminación atmosférica, pero puede no interesar a una cadena de supermercados. Por otro lado, responder a la pregunta de si las ventas de pepperoni son mayores cuando se exhibe junto a la salsa y la masa de la pizza o cuando se exhibe con las otras carnes envasadas sería de interés para una cadena de tiendas de comestibles, pero no para personas de otras industrias.

También hay que comprobar que la pregunta **no haya sido ya respondida.** Con la reciente explosión de datos, la creciente cantidad de datos disponibles públicamente y la aparentemente interminable literatura científica y otros recursos, no es raro descubrir que su pregunta de interés ya ha sido respondida. Algunas investigaciones y discusiones con expertos pueden ayudar a resolver esto, y también pueden ser útiles porque incluso si la pregunta específica que usted tiene en mente no ha sido contestada, las preguntas relacionadas pueden haber sido contestadas y las respuestas a estas preguntas relacionadas son informativas para decidir si o cómo usted procede con su pregunta específica.

La pregunta también debe partir de un marco **plausible**. En otras palabras, la pregunta anterior sobre la relación entre las ventas de salchichón y su ubicación en la tienda es plausible porque los compradores que adquieren ingredientes de pizza tienen más probabilidades que otros compradores de estar interesados en el salchichón y es más probable que lo compren si lo ven al mismo tiempo que seleccionan los demás ingredientes de la pizza. Una pregunta menos plausible sería si las ventas de pepperoni se correlacionan con las de yogur, a no ser que se tenga algún conocimiento previo que sugiera que deberían estar correlacionadas.

Si se formula una pregunta cuyo marco no es plausible, es probable que se obtenga una respuesta difícil de interpretar o en la que se pueda confiar. En el caso de la pregunta sobre el pepperoni y el yogur, si se descubre que están correlacionados, se plantean muchas preguntas sobre el propio resultado: ¿es realmente correcto?, ¿por qué están correlacionados?, ¿hay otra explicación?, y otras. Puedes asegurarte de que tu pregunta se basa en un marco plausible utilizando tus propios conocimientos sobre el tema e investigando un poco, lo que, en conjunto, puede ayudarte a resolver si tu pregunta se basa en un marco plausible.

Por supuesto, la pregunta también debe tener **respuesta**. Aunque tal vez no sea necesario decirlo, vale la pena señalar que algunas de las mejores preguntas no tienen respuesta, ya sea porque los datos no existen o porque no hay forma de recopilarlos por falta de recursos, viabilidad o problemas éticos. Por ejemplo, es bastante plausible que haya defectos en el funcionamiento de ciertas células del cerebro que causen el autismo, pero no es posible realizar biopsias cerebrales para recoger células vivas para estudiar, lo que sería necesario para responder a esta pregunta.

La especificidad es también una característica importante de una buena pregunta. Un ejemplo de pregunta general es: ¿Le conviene llevar una dieta más sana? Trabajar para conseguir la especificidad afinará la pregunta e informará directamente de los pasos que hay que dar cuando se empiece a estudiar los datos. Una pregunta más específica surge después de preguntarse a qué se refiere con una dieta "más sana" y cuándo dice que algo es "mejor para usted". El proceso de aumentar la especificidad debería conducir a una pregunta final más refinada, como por ejemplo "¿Consumir al menos 5 raciones al día de frutas y verduras frescas provoca menos infecciones del tracto respiratorio superior (resfriados)?". Con este grado de especificidad, su plan de ataque es mucho más claro y la respuesta que obtendrá al final del análisis de los datos será

más interpretable, ya que recomendará o no recomendará la acción específica de comer al menos 5 raciones de frutas y verduras frescas al día como medio de protección contra las infecciones de las vías respiratorias superiores.

3.4 Traducir una pregunta en un problema de datos

Otro aspecto que hay que tener en cuenta a la hora de elaborar la pregunta es lo que ocurrirá cuando se traduzca en un problema de datos. Toda pregunta debe operacionalizarse como un análisis de datos que conduzca a un resultado. Detenerse a pensar cómo serían los resultados del análisis de datos y cómo podrían interpretarse es importante, ya que puede evitar que se pierda mucho tiempo embarcándose en un análisis cuyo resultado no es interpretable. Aunque a lo largo del libro se expondrán muchos ejemplos de preguntas que conducen a resultados interpretables y significativos, lo más fácil es empezar por pensar qué tipo de preguntas *no* conducen a respuestas interpretables.

El tipo típico de pregunta que no cumple este criterio es una pregunta que utiliza datos inadecuados. Por ejemplo, su pregunta puede ser si tomar un suplemento de vitamina D se asocia con menos dolores de cabeza, y usted planea responder a esa pregunta utilizando el número de veces que una persona tomó un analgésico como marcador del número de dolores de cabeza que tuvo. Es posible que encuentre una asociación entre la toma de suplementos de vitamina D y la toma de menos analgésicos, pero no estará claro cuál es la interpretación de este resultado. De hecho, es posible que las personas que toman suplementos de vitamina D también tiendan a ser menos propensas a tomar otros medicamentos de venta libre sólo porque "evitan la medicación", y no porque realmente

menos dolores de cabeza. También puede ser que utilicen menos analgésicos porque tienen menos dolor en las articulaciones u otros tipos de dolor, pero no menos dolores de cabeza. Otra interpretación, por supuesto, es que sí tienen menos dolores de cabeza, pero el problema es que no se puede determinar si ésta es la interpretación correcta o si una de las otras interpretaciones es la correcta. En esencia, el problema de esta pregunta es que, para una única respuesta posible, hay múltiples interpretaciones. Este escenario de múltiples interpretaciones surge cuando al menos una de las variables que utiliza (en este caso, el uso de analgésicos) no es una buena medida del concepto que realmente busca (en este caso, los dolores de cabeza). Para evitar este problema, deberá asegurarse de que los datos disponibles para responder a su pregunta proporcionan medidas razonablemente específicas de los factores necesarios para responder a su pregunta.

Un problema relacionado que interfiere con la interpretación de los resultados es la confusión. La confusión es un problema potencial cuando su pregunta se refiere a la relación entre factores, como la toma de vitamina D y la frecuencia de los dolores de cabeza. Una breve descripción del concepto de confusión es que está presente cuando un factor que no estaba necesariamente considerando en su pregunta está relacionado tanto con su exposición de interés (en el ejemplo, tomar suplementos de vitamina D) como con su resultado de interés (tomar medicación analgésica). Por ejemplo, los ingresos podrían ser un con- flicto, porque pueden estar relacionados tanto con la toma de suplementos de vitamina D como con la frecuencia de los dolores de cabeza, ya que las personas con mayores ingresos pueden tender a tomar un suplemento y a tener menos problemas de salud crónicos, como los dolores de cabeza. Por lo general, siempre que disponga de datos sobre los ingresos, podrá ajustar este con- flicto y reducir el número de posibles interpretaciones de la respuesta a su pregunta. A medida que vaya afinando su pregunta

dedique algún tiempo a identificar los posibles factores de confusión y a pensar si su conjunto de datos incluye información sobre estos posibles factores de confusión.

Otro tipo de problema que puede surgir cuando se utilizan datos inadecuados es que el resultado no es interpretable porque la forma subyacente en que se recogieron los datos conduce a un resultado sesgado. Por ejemplo, imagine que utiliza un conjunto de datos creado a partir de una encuesta a mujeres que han tenido hijos. La encuesta incluye información sobre si sus hijos tenían autismo y si declararon haber comido sushi durante el embarazo, y usted ve una asociación entre el informe de comer sushi durante el embarazo y tener un hijo con autismo. Sin embargo, debido a que las mujeres que han tenido un hijo con una condición de salud recuerdan las exposiciones, como el pescado crudo, que se produjeron durante el embarazo de manera diferente a las que han tenido hijos sanos, la asociación observada entre la exposición al sushi y el autismo puede ser sólo la manifestación de la tendencia de la madre a centrarse más en los eventos durante el embarazo cuando tiene un hijo con una condición de salud. Este es un ejemplo de sesgo de recuerdo, pero hay muchos tipos de sesgo que pueden ocurrir.

El otro sesgo importante que hay que entender y tener en cuenta a la hora de refinar la pregunta es el sesgo de selección, que se produce cuando los datos que se analizan se han recogido de forma que se infla la proporción de personas que tienen ambas características por encima de lo que existe en la población general. Si en un estudio se anuncia que se trata de un estudio sobre el autismo y la dieta durante el embarazo, es muy posible que las mujeres que comen pescado crudo y tienen un hijo con autismo tengan más probabilidades de responder a la encuesta que las que tienen una de estas condiciones o ninguna de ellas. Este escenario llevaría a una respuesta sesgada a su pregunta sobre la ingesta de sushi de las madres durante el embarazo y el riesgo de autismo en sus hijos. Una buena regla general es que si

Si está examinando las relaciones entre dos factores, el sesgo puede ser un problema si tiene más (o menos) probabilidades de observar individuos con ambos factores debido a la forma en que se seleccionó la población o a la forma en que una persona puede recordar el pasado al responder a una encuesta. Se hablará más sobre el sesgo en los capítulos siguientes (Inferencia: una cartilla e Interpretación de los resultados), pero el mejor momento para considerar sus efectos en el análisis de los datos es cuando se identifica la pregunta a la que se va a responder y se piensa en cómo se va a responder a la pregunta con los datos disponibles.

3.5 Estudio de caso

Joe trabaja para una empresa que fabrica diversos dispositivos y aplicaciones de seguimiento del estado físico y cuyo nombre es Fit on Fleek. El objetivo de Fit on Fleek es, como el de muchas empresas tecnológicas de nueva creación, utilizar los datos que recogen de los usuarios de sus dispositivos para hacer un marketing dirigido a diversos productos. El producto que les gustaría comercializar es uno nuevo que acaban de desarrollar y que aún no han empezado a vender, que es un rastreador del sueño y una aplicación que hace un seguimiento de las distintas fases del sueño, como el sueño REM, y también ofrece consejos para mejorar el sueño. El rastreador del sueño se llama Sleep on Fleek.

El jefe de Joe le pide que analice los datos que la empresa tiene sobre los usuarios de sus dispositivos y aplicaciones de seguimiento de la salud para identificar a los usuarios de los anuncios dirigidos a Sleep on Fleek. Fit on Fleek dispone de los siguientes datos de cada uno de sus clientes: información demográfica básica, número de pasos caminados al día, número de tramos de escaleras subidos al día, horas de sedentarismo al día, horas de alerta al día, horas de somnolencia al día y horas de sueño al día (pero no información más detallada sobre el sueño que el

rastreador del sueño rastrearía).

Aunque Joe tiene un objetivo en mente, extraído de una

Si bien es cierto que Joe ha hablado con su jefe y sabe qué tipos de datos están disponibles en la base de datos Fit on Fleek, todavía no tiene una pregunta. Este escenario, en el que a Joe se le da un objetivo, pero no una pregunta, es común, por lo que la primera tarea de Joe es traducir el objetivo en una pregunta, y esto requerirá alguna comunicación de ida y vuelta con su jefe. El enfoque de las comunicaciones informales que tienen lugar durante el proceso del proyecto de análisis de datos, se trata en detalle en el capítulo de Comunicación. Tras unas cuantas charlas, Joe se decide por la siguiente pregunta: "¿Qué usuarios de Fit on Fleek no duermen lo suficiente?". Él y su jefe están de acuerdo en que los clientes que probablemente estarían más interesados en comprar el dispositivo y la aplicación Sleep on Fleek son aquellos que parecen tener problemas con el sueño, y el problema más fácil de rastrear y probablemente el más común es no dormir lo suficiente.

Se podría pensar que, puesto que Joe ya tiene una pregunta, debería pasar a descargar los datos y empezar a hacer análisis ex- ploratorios, pero hay un poco de trabajo que Joe todavía tiene que hacer para refinar la pregunta. Las dos tareas principales que Joe debe abordar son (1) pensar en cómo su pregunta cumple, o no, las características de una buena pregunta y
(2) para determinar qué tipo de pregunta está formulando, de modo que comprenda bien qué tipo de conclusiones pueden (y no pueden) extraerse cuando haya terminado el análisis de los datos.

Joe repasa las características de una buena pregunta y sus expectativas son que su pregunta tenga todas estas características: -de interés -que no haya sido ya respondida -fundada en un marco plausible -contestable -específica

La respuesta que obtendrá al final de su análisis (cuando traduzca su pregunta en un problema de datos) también debe ser interpretable.

A continuación, piensa en lo que sabe sobre la pregunta y, a su juicio, la pregunta es de interés, ya que su jefe expresó su interés.

También sabe que la pregunta no puede haber sido analizada ya, ya que su jefe le indicó que no lo había hecho y una revisión de los análisis de datos anteriores de la empresa revela que no hay ningún análisis anterior diseñado para responder a la pregunta.

A continuación, evalúa si la pregunta se basa en un marco plausible. La pregunta "¿Qué usuarios de Fit on Fleek no duermen lo suficiente?" parece estar basada en un marco plausible, ya que tiene sentido que las personas que duermen poco estén interesadas en tratar de mejorar su sueño haciendo un seguimiento del mismo. Sin embargo, Joe se pregunta si la duración del sueño es el mejor marcador para saber si una persona siente que está durmiendo poco. Conoce a algunas personas que duermen regularmente poco más de 5 horas por noche y parecen estar satisfechas con su sueño. Joe acude a un especialista en medicina del sueño y se entera de que una mejor medida para saber si alguien está afectado por la falta de sueño o por un sueño de mala calidad es la somnolencia diurna. Resulta que su expectativa inicial de que la pregunta se basaba en un marco plausible no coincidía con la información que recibió cuando habló con un experto en contenido. Así que revisa su pregunta para que se ajuste a sus expectativas de verosimilitud y la pregunta revisada es: ¿Qué usuarios de Fit on Fleek tienen somnolencia durante el día?

Joe se detiene para asegurarse de que esta pregunta es, de hecho, anerable con los datos que tiene a su disposición, y confirma que lo es. También se detiene a pensar en la especificidad de la pregunta. Cree que es específica, pero hace el ejercicio de debatir la pregunta con sus colegas para recabar información sobre la especificidad de la misma. Cuando se plantea la idea de responder a esta pregunta, su col-

Las ligas le hacen muchas preguntas sobre lo que significan varias partes de la pregunta: ¿qué significa "qué usuarios"? ¿significa esto: ¿Cuáles son las características demográficas de los usuarios que tienen somnolencia? ¿O algo más? ¿Y la "somnolencia durante el día"? ¿Debe esta frase significar cualquier somnolencia en cualquier día? ¿O la somnolencia que dura al menos una cierta cantidad de tiempo en al menos un cierto número de días? La conversación con los colegas fue muy informativa e indicó que la pregunta no era muy específica. Joe revisa su pregunta para que ahora sea específica: "¿Qué características demográficas y de salud identifican a los usuarios que tienen más probabilidades de sufrir somnolencia crónica, definida como al menos un episodio de somnolencia al menos cada dos días?"

Joe pasa ahora a pensar en cuáles son las posibles respuestas a sus preguntas y si serán interpretables. Joe identifica dos posibles resultados de su análisis: (1) no hay características que identifiquen a las personas que tienen somnolencia diurna crónica o (2) hay una o más características que identifican a las personas con somnolencia diurna crónica. Estas dos posibilidades son interpretables y significativas. Para la primera, Joe llegaría a la conclusión de que no sería posible dirigir los anuncios del rastreador Sleep on Fleek a las personas que se prevé que tienen somnolencia diurna crónica, y para la segunda, llegaría a la conclusión de que es posible dirigir el anuncio, y sabría qué característica(s) utilizar para seleccionar a las personas para los anuncios dirigidos.

Ahora que Joe tiene una buena pregunta, después de repetir los 3 pasos del epiciclo mientras consideraba si su pregunta cumplía con cada una de las características de una buena pregunta, el siguiente paso es averiguar qué tipo de pregunta tiene. Para ello, sigue un proceso de reflexión similar al que utilizó para cada una de las características anteriores. Empieza a pensar que su pregunta es exploratoria, pero al revisar la descripción y los ejemplos de

una pregunta exploratoria, se da cuenta de que, aunque algunas partes del análisis que hará para responder a la pregunta serán exploratorias, en última instancia su pregunta es más que exploratoria porque su respuesta predecirá qué usuarios son propensos a tener somnolencia diurna crónica, por lo que su pregunta es una pregunta de predicción. Identificar el tipo de pregunta es muy útil porque, junto con una buena pregunta, ahora sabe que tiene que utilizar un enfoque de predicción en sus análisis, en particular en la fase de construcción del modelo (véase el capítulo sobre modelización formal).

3.6 Reflexiones finales

A estas alturas, debería estar preparado para aplicar los 3 pasos del epiciclo a la hora de plantear y refinar una pregunta. Si es un analista de datos experimentado, gran parte de este proceso puede ser automático, por lo que es posible que no sea del todo consciente de algunas partes del proceso que le conducen a una buena pregunta. Hasta que llegue a este punto, este capítulo puede servirle como recurso útil cuando se enfrente a la tarea de desarrollar una buena pregunta. En los próximos capítulos, hablaremos de qué hacer con los datos ahora que tienes una buena pregunta en la mano.

4. Análisis exploratorio de datos

El análisis exploratorio de datos es el proceso de exploración de los datos, y suele incluir el examen de la estructura y los componentes del conjunto de datos, las distribuciones de las variables individuales y las relaciones entre dos o más variables. La herramienta más utilizada para el análisis exploratorio de datos es la visualización de los datos mediante una representación gráfica de los mismos. La visualización de los datos es, sin duda, la herramienta más importante para el análisis exploratorio de datos, porque la información transmitida por la visualización gráfica puede absorberse muy rápidamente y porque, por lo general, es fácil reconocer patrones en una visualización gráfica.

Hay varios objetivos del análisis exploratorio de datos, que son:

1. Para determinar si hay algún problema con su conjunto de datos.
2. Determinar si la pregunta que se hace puede ser respondida por los datos que se tienen.
3. Desarrollar un esbozo de la respuesta a su pregunta.

Su aplicación del análisis exploratorio de datos estará guiada por su pregunta. La pregunta de ejemplo utilizada en este capítulo es: "¿Los condados del este de Estados Unidos tienen niveles de ozono más altos que los condados del oeste de Estados Unidos?" En este caso, usted explorará los datos para determinar si hay problemas con el conjunto de datos, y para determinar si puede responder a su pregunta con este conjunto de datos.

Para responder a la pregunta, por supuesto, se necesitan datos sobre el ozono, el condado y la región de EEUU. El siguiente paso es utilizar el análisis exploratorio de datos para empezar a responder a tu pregunta, lo que podría incluir la visualización de gráficos de caja del ozono por región de los Estados Unidos. Al final del análisis exploratorio de datos, deberías tener una buena idea de cuál es la respuesta a tu pregunta y estar armado con suficiente información para pasar a los siguientes pasos del análisis de datos.

Es importante señalar que aquí, de nuevo, se aplica el concepto de epiciclo de análisis. Debe tener una expectativa de cómo será su conjunto de datos y si su pregunta puede ser respondida por los datos que tiene. Si el contenido y la estructura del conjunto de datos no coinciden con sus expectativas, tendrá que volver atrás y averiguar si su expectativa era correcta (pero había un problema con los datos) o, alternativamente, su expectativa era incorrecta, por lo que no puede utilizar el conjunto de datos para responder a la pregunta y tendrá que encontrar otro conjunto de datos.

También deberías tener alguna expectativa sobre cuáles serán los niveles de ozono, así como si el ozono de una región debería ser mayor (o menor) que el de otra. Al pasar al paso 3 de empezar a responder a tu pregunta, volverás a aplicar el epiciclo de análisis, de modo que si, por ejemplo, los niveles de ozono en el conjunto de datos son más bajos de lo que esperabas al mirar los datos publicados anteriormente, tendrás que hacer una pausa y averiguar si hay un problema con tus datos o si tu expectativa era incorrecta. Su expectativa podría ser incorrecta, por ejemplo, si su fuente de información para establecer su expectativa sobre los niveles de ozono eran datos recogidos hace 20 años (cuando los niveles eran probablemente más altos) o de una sola ciudad en los EE.UU. Entraremos en más detalle con el estudio de caso a continuación, pero esto debería darle una visión general sobre el enfoque y los objetivos del análisis exploratorio de datos.

4.1 Lista de verificación del análisis exploratorio de datos: Un estudio de caso

En esta sección repasaremos una "lista de comprobación" informal de las cosas que hay que hacer al embarcarse en un análisis exploratorio de datos. Como ejemplo, utilizaré un conjunto de datos sobre los niveles horarios de ozono en Estados Unidos para el año 2014. Los elementos de la lista de comprobación son

1. Formule su pregunta
2. Lea sus datos
3. Compruebe el embalaje
4. Mira la parte superior e inferior de tus datos
5. Comprueba tus "n"
6. Validar con al menos una fuente de datos externa
7. Hacer una parcela
8. Pruebe primero la solución fácil
9. Seguimiento

A lo largo de este ejemplo representaremos un análisis en curso con código R y datos reales. Algunos de los ejemplos y recomendaciones serán específicos del entorno de análisis estadístico R, pero la mayoría deberían ser aplicables a cualquier sistema de software. No es necesario dominar R para entender las ideas principales del ejemplo. Siéntase libre de saltarse las secciones de código.

4.2 Formule su pregunta

Ya hemos hablado en este libro de la importancia de formular correctamente una pregunta. La formulación de una pregunta puede ser una forma útil de guiar el análisis exploratorio de datos

y limitar el número exponencial de caminos que se pueden tomar con cualquier conjunto de datos de tamaño considerable. En particular, una pregunta o una hipótesis *definida* puede servir como herramienta de reducción de la dimensión que puede eliminar las variables que no son inmediatamente relevantes para la pregunta.

Por ejemplo, en este capítulo estudiaremos un conjunto de datos sobre contaminación atmosférica de la Agencia de Protección del Medio Ambiente (EPA) de Estados Unidos. Una pregunta general que se podría hacer es

> ¿Son los niveles de contaminación del aire más altos en la costa este que en la costa oeste?

Pero una pregunta más específica podría ser

> ¿Son los niveles de ozono por hora más altos en Nueva York que en Los Ángeles?

Hay que tener en cuenta que ambas preguntas pueden ser interesantes, y que ninguna es correcta o incorrecta. Pero la primera pregunta requiere analizar todos los contaminantes en toda la costa este y oeste, mientras que la segunda pregunta sólo requiere analizar un único contaminante en dos ciudades.

Suele ser una buena idea dedicar unos minutos a averiguar cuál es la pregunta que *realmente le interesa*, y reducirla para que sea lo más específica posible (sin dejar de ser interesante).

Para este capítulo, consideraremos la siguiente cuestión:

> ¿Los condados del este de Estados Unidos tienen niveles de ozono más altos que los del oeste?

Como nota al margen, una de las preguntas más importantes que se pueden responder con un análisis exploratorio de datos es "¿Tengo los datos adecuados para responder a esta pregunta?". A menudo, esta pregunta es difícil de responder al principio, pero puede aclararse a medida que ordenamos y examinamos los datos.

4.3 Lea sus datos

La siguiente tarea en cualquier análisis exploratorio de datos es leer algunos datos. A veces los datos vendrán en un formato muy desordenado y tendrá que hacer alguna limpieza. Otras veces, otra persona habrá limpiado los datos por ti, de modo que te ahorrarás el dolor de tener que hacer la limpieza.

No vamos a pasar por el dolor de la limpieza de un conjunto de datos aquí, no porque no sea importante, sino más bien porque a menudo no hay mucho conocimiento generalizable para obtener de ir a través de él. Cada conjunto de datos tiene sus peculiaridades, por lo que, por ahora, probablemente sea mejor no enfrascarse en los detalles.

Aquí tenemos un conjunto de datos relativamente limpio de la EPA de Estados Unidos sobre las mediciones horarias de ozono en todo el país para el año 2014. Los datos están disponibles en la página web del Sistema de Calidad del Aire de la EPA[1]. Simplemente he descargado el archivo zip de la página web, he descomprimido el archivo y he puesto el archivo resultante en un directorio llamado "data". Si quieres ejecutar este código tendrás que utilizar la misma estructura de directorios.

El conjunto de datos es un archivo de valores separados por comas (CSV), donde cada fila del archivo contiene una medición horaria de ozono en algún lugar del país.

NOTA: Ejecutar el código siguiente puede llevar unos minutos. Hay 7.147.884 filas en el archivo CSV. Si tarda mucho,

[1] http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

puede leer un subconjunto especificando un valor para `n_max` a `read_csv()` que sea mayor que 0.

```
> biblioteca(readr)
> ozono <- read_csv("data/hourly_44201_2014.csv",
+col_types          = "ccccinnccccncnncccc")
```

El paquete `readr` de Hadley Wickham es un buen paquete para leer archivos planos (como los archivos CSV) *muy* rápido, o al menos mucho más rápido que las funciones incorporadas de R. Hace algunas concesiones para obtener esa velocidad, por lo que estas funciones no son siempre apropiadas, pero sirven para nuestros propósitos aquí.

La cadena de caracteres proporcionada al argumento `col_types` especifica la clase de cada columna en el conjunto de datos. Cada letra representa la clase de una columna: "c" para carácter, "n" para numérico", e "i" para entero. No, no sabía por arte de magia las clases de cada columna; simplemente miré rápidamente el archivo para ver cuáles eran las clases de las columnas. Si hay demasiadas columnas, no puedes especificar `col_types` y `read_csv()` intentará averiguarlo por ti.

Por comodidad, podemos reescribir los nombres de las columnas para eliminar los espacios.

```
> names(ozono) <- make.names(names(ozono))
```

4.4 Compruebe el embalaje

¿Alguna vez has recibido un regalo *antes de* la hora en la que podías abrirlo? Seguro que a todos nos ha pasado. El problema es que el regalo está envuelto, pero quieres saber desesperadamente qué hay dentro. ¿Qué puede hacer una persona en esas circunstancias? Bueno, puede agitar la caja un poco, tal vez golpearla con el nudillo para ver si hace un sonido hueco, o incluso

pesarlo para ver su peso. Así es como debes pensar en tu conjunto de datos antes de empezar a analizarlo de verdad.

Suponiendo que no reciba ninguna advertencia o error al leer el conjunto de datos, ahora debería tener un objeto en su espacio de trabajo llamado `ozono`. Suele ser una buena idea hurgar un poco en ese objeto antes de abrir el papel de regalo.

Por ejemplo, debe comprobar el número de filas

```
> nrow(ozono)
[1] 7147884
```

y columnas.

```
> ncol(ozono)
[1] 23
```

¿Recuerdas cuando dijimos que había 7.147.884 filas en el archivo? ¿Cómo coincide con lo que hemos leído? Este conjunto de datos también tiene relativamente pocas columnas, por lo que podría comprobar el archivo de texto original para ver si el número de columnas impresas (23) aquí coincide con el número de columnas que se ve en el archivo original.

Otra cosa que puede hacer en R es ejecutar `str()` en el conjunto de datos. Esta suele ser una operación segura en el sentido de que, incluso con un conjunto de datos muy grande, la ejecución de `str()` no debería llevar demasiado tiempo.

```
> str(ozono)
Clases 'tbl_df', 'tbl' y 'data.frame'            :7147884obs. de 23 variab\
les:
 $ State.Code          : chr "01" "01" "01" "01" ...
 $ County.Code         : chr "003" "003" "003" "003" ...
 $ Site.Num            : chr "0010" "0010" "0010" "0010" ...
 $ Parámetro.Código    : chr "44201" "44201" "44201" "44201" ...
 $ POC                 : int 1 1 1 1 1 1 1 1 1 ...
 $ Latitud             : num 30,5 30,5 30,5 30,5 30,5 ...
 $ Longitud            : num -87,9 -87,9 -87,9 -87,9 -87,9 ...
 $ Datum               : chr "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parámetro.Nombre    : chr "Ozono" "Ozono" "Ozono" "Ozono" ...
 $ Date.Local          : chr "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
 $ Time.Local          : chr "01:00" "02:00" "03:00" "04:00" ...
 $ Date.GMT            : chr "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
 $ Time.GMT            : chr "07:00" "08:00" "09:00" "10:00" ...
 Muestra.Medida : num 0.047 0.047 0.043 0.038 0.035 0.035 0.0\
34 0.037 0.044 0.046 ...
 $ Unidades.de.Medida : chr "Partes por millón" "Partes por
 millón"\N-"Partes por millón"...
 $ MDL                 : num 0,005 0,005 0,005 0,005 0,005 0,005
05 0.005 0.005 0.005 ...
 $ Incertidumbre : num NA NA NA NA NA NA NA ...
 $ Calificador         : chr "" "" "" "" ...
 $ Method.Type : chr "FEM" "FEM" "FEM" "FEM" ...
 $ Method.Name : chr "INSTRUMENTAL - ULTRA VIOLETA" "INSTRUME NTAL -
ULTRA VIOLETA" "INSTRUMENTAL - ULTRA VIOLETA" "INSTRUMENTAL - U\ LTRA
VIOLETA" ...
 $ Estado.Nombre       : chr    "Alabama" "Alabama" "Alabama" "Alabama"
 ...
 Nombre del            : chr              "Baldwin" "Baldwin" "Baldwin"
 condado                                                  "Baldwin"\N-
 ...
 $ Fecha.de.la.última.modificación: chr "2014-06-30" "2014-06-30" "2014-06-30" \N
"2014-06-30" ...
```

La salida de `str()` duplica alguna información que ya tenemos, como el número de filas y columnas. Lo que es más importante, puede examinar las *clases* de cada una de las columnas para asegurarse de que están correctamente especificadas (es decir, num-

son `numéricas` y las cadenas son `de carácter`, etc.). Como hemos especificado previamente todas las clases de columnas en `read_csv()`, todas deberían coincidir con lo que hemos especificado.

A menudo, con estas sencillas maniobras, se pueden identificar posibles problemas con los datos antes de lanzarse de cabeza a un complicado análisis de datos.

4.5 Mira la parte superior e inferior de tus datos

A menudo es útil mirar el "principio" y el "final" de un conjunto de datos justo después de comprobar el embalaje. Esto le permite saber si los datos se leyeron correctamente, si están bien formateados y si todo está allí. Si sus datos son de series temporales, asegúrese de que las fechas al principio y al final del conjunto de datos coinciden con lo que usted espera que sea el período de tiempo inicial y final.

En R, se puede echar un vistazo a la parte superior e inferior de los datos con las funciones `head()` y `tail()`.

Aquí está la parte superior.

```
> head(ozono[, c(6:7, 10)])
  Latitud Longitud Fecha.Local
    130.498-87.88141 2014-03-01
    230.498-87.88141 2014-03-01
    330.498-87.88141 2014-03-01
    430.498-87.88141 2014-03-01
    530.498-87.88141 2014-03-01
    630.498-87.88141 2014-03-01
```

Para ser breve, sólo he tomado algunas columnas. Y aquí está el fondo.

```
> tail(ozono[, c(6:7, 10)])
        Latitud Longitud Fecha.Local
7147879 18.17794 -65.91548 2014-09-30
7147880 18.17794 -65.91548 2014-09-30
7147881 18.17794 -65.91548 2014-09-30
7147882 18.17794 -65.91548 2014-09-30
7147883 18.17794 -65.91548 2014-09-30
7147884 18.17794 -65.91548 2014-09-30
```

La función `tail()` puede ser especialmente útil porque a menudo habrá algún problema al leer el final de un conjunto de datos y si no lo compruebas específicamente nunca lo sabrás. A veces hay un formato extraño al final o algunas líneas de comentario ex- tra que alguien decidió pegar al final. Esto es particularmente común con los datos que se exportan de las hojas de cálculo de Microsoft Excel.

Asegúrese de comprobar todas las columnas y verificar que todos los datos de cada columna tienen el aspecto que se supone que deben tener. Esto no es un enfoque infalible, porque sólo estamos viendo unas pocas filas, pero es un buen comienzo.

4.6 ABC: Siempre revisa tus "n"

En general, contar cosas suele ser una buena manera de averiguar si algo va mal o no. En el caso más sencillo, si esperas que haya 1.000 observaciones y resulta que sólo hay 20, sabes que algo debe haber ido mal en alguna parte. Pero hay otras áreas que puedes comprobar dependiendo de tu aplicación. Para hacerlo correctamente, tienes que identificar algunos *puntos de referencia que se* pueden utilizar para comprobar tus datos. Por ejemplo, si estás recogiendo datos sobre personas, como en una encuesta o un ensayo clínico, debes saber cuántas personas hay en tu estudio. Eso es algo que debes comprobar en tu conjunto de datos, para hacer

Asegúrate de que tienes datos de todas las personas de las que pensabas que ibas a tener datos.

En este ejemplo, utilizaremos el hecho de que el conjunto de datos contiene puramente datos *horarios* para *todo el país*. Estos serán nuestros dos puntos de referencia para la comparación.

Aquí tenemos datos de ozono por hora que provienen de monitores de todo el país. Los monitores deberían estar monitoreando continuamente durante el día, por lo que todas las horas deberían estar representadas. Podemos echar un vistazo a la variable `Time.Local` para ver a qué hora se registran las mediciones.

```
> head(tabla(ozono$Hora.Local))

 00:00 00:01 01:00 01:02 02:00 02:03
            2886982       2908712       2837092
```

Una cosa que observamos aquí es que, aunque casi todas las mediciones del conjunto de datos se registran como tomadas a la hora, algunas se toman a horas ligeramente diferentes. El número de lecturas que se toman a estas horas es tan pequeño que no nos importa. Pero parece un poco extraño, así que vale la pena hacer una comprobación rápida.

Podemos ver qué observaciones se midieron a la hora "00:01".

```
> library(dplyr)
> filter(ozono, Hora.Local == "13:14") %>%
+select   (State.Name, County.Name, Date.Local,
+Time             .Local, Sample.Measurement)
Fuente: marco de datos local [2 x 5]

Nombre.del.estado Nombre.del.condado Fecha.local Hora.local
       (chr)       (chr)      (chr)      (chr)
1      Nueva     Franklin 2014-09-30      13:14
        York
2      Nueva     Franklin 2014-09-30      13:14
        York
 Variables no mostradas: Muestra.Medida (dbl)
```

Podemos ver que se trata de un monitor en el condado de Franklin, Nueva York, y que las mediciones se realizaron el 30 de septiembre de 2014. Qué pasaría si simplemente sacáramos todas las mediciones tomadas en este monitor en esta fecha?

```
> filter(ozono, Estado.Código == "36"
         +&County.Code == "033"
         +&Date.Local == "2014-09-30") %>%
+select   (Date.Local, Time.Local,
+Muestra         .Medida) %>%
+as       .data.frame
   Fecha.Local Hora.Local Muestra.Medida 1
                  2014-09-3000:01    0.011
2                                    2014-09-3001:020.012
3                                    2014-09-3002:030.012
4                                    2014-09-3003:040.011
5                                    2014-09-3004:050.011
6                                    2014-09-3005:060.011
7                                    2014-09-3006:070.010
8                                    2014-09-3007:080.010
9                                    2014-09-3008:090.010
10                                   2014-09-3009:100.010
11                                   2014-09-3010:110.010
12                                   2014-09-3011:120.012
13                                   2014-09-3012:130.011
14                                   2014-09-3013:140.013
15                                   2014-09-3014:150.016
16                                   2014-09-3015:160.017
17                                   2014-09-3016:170.017
```

```
18 2014-09-30      17:18             0.015
19 2014-09-30      18:19             0.017
20 2014-09-30      19:20             0.014
21 2014-09-30      20:21             0.014
22 2014-09-30      21:22             0.011
23 2014-09-30      22:23             0.010
24 2014-09-30      23:24             0.010
25 2014-09-30      00:01             0.010
26 2014-09-30      01:02             0.011
27 2014-09-30      02:03             0.011
28 2014-09-30      03:04             0.010
29 2014-09-30      04:05             0.010
30 2014-09-30      05:06             0.010
31 2014-09-30      06:07             0.009
32 2014-09-30      07:08             0.008
33 2014-09-30      08:09             0.009
34 2014-09-30      09:10             0.009
35 2014-09-30      10:11             0.009
36 2014-09-30      11:12             0.011
37 2014-09-30      12:13             0.010
38 2014-09-30      13:14             0.012
39 2014-09-30      14:15             0.015
40 2014-09-30      15:16             0.016
41 2014-09-30      16:17             0.016
42 2014-09-30      17:18             0.014
43 2014-09-30      18:19             0.016
44 2014-09-30      19:20             0.013
45 2014-09-30      20:21             0.013
46 2014-09-30      21:22             0.010
47 2014-09-30      22:23             0.009
48 2014-09-30      23:24             0.009
```

Ahora podemos ver que este monitor sólo registra sus valores a horas impares, en lugar de a horas. Al parecer, viendo la salida anterior, este es el único monitor del país que hace esto, así que probablemente no es algo de lo que debamos preocuparnos.

Dado que la EPA controla la contaminación en todo el país, debería haber una buena representación de los estados. Quizás deberíamos ver exactamente cuántos estados están representados en esta

conjunto de datos.

```
> select(ozono, Estado.Nombre) %>% unique %>%
nrow [1] 52
```

Parece que la representación es demasiado buena: hay 52 estados en el conjunto de datos, ¡pero sólo 50 estados en Estados Unidos!

Podemos echar un vistazo a los elementos únicos del `Estado.Nombre` variable para ver lo que está pasando.

```
> unique(ozono$Estado.Nombre)
 [1]                        "Alabama""Alaska"
 [3]                        "Arizona""Arkansas"
 [5]                        "California""Colorado"
 [7]                        "Connecticut""Delaware"
 [9] "Distrito de Columbia" "Florida"
[11]                        "Georgia""Hawaii"
[13]                        "Idaho""Illinois"
[15]                        "Indiana""Iowa"
[17]                        "Kansas""Kentucky"
[19]                        "Luisiana""Maine"
[21]                        "Maryland""Massachusetts"
[23]                        "Michigan""Minnesota"
[25]                        "Mississippi""Missouri"
[27]                        "Montana""Nebraska"
[29] "                      Nevada""NewHampshire"
[31] "Nueva                 Jersey""NuevoMéxico"
[33] "Nueva                 York""Carolina     delNorte"
[35] "                      Dakotadel Norte    ""Ohio"
[37]                        "Oklahoma""Oregón"
[39]                        "Pennsylvania""RhodeIsland"
[41] "                      Carolinadel        Sur""Dakota      delSur"
[43]                        "Tennessee""Texas"
[45]                        "Utah""Vermont"
[47]                        "Virginia""Washington"
[49] "West                  Virginia""Wisconsin"
[51]                        "Wyoming""PuertoRico"
```

Ahora podemos ver que Washington, D.C. (Distrito de Columbia) y Puerto Rico son los estados "extra" incluidos en el conjunto de datos.

Como es evidente que forman parte de los EE.UU. (pero no son estados oficiales de la unión) todo parece correcto.

Para este último análisis se utilizó algo que comentaremos en la siguiente sección: datos externos. Sabíamos que sólo hay 50 estados en EE.UU., así que ver los nombres de 52 estados fue un desencadenante inmediato de que algo podría estar mal. En este caso, todo estaba bien, pero validar tus datos con una fuente de datos externa puede ser muy útil. Lo que nos lleva a ....

4.7 Validación con al menos una fuente de datos externa

Asegurarse de que los datos coinciden con algo fuera del conjunto de datos es muy importante. Le permite asegurarse de que las mediciones se ajustan aproximadamente a lo que deberían ser y sirve para comprobar qué *otras* cosas podrían estar mal en su conjunto de datos. La validación externa a menudo puede ser tan simple como comprobar los datos con un solo número, como haremos aquí.

En Estados Unidos tenemos normas nacionales de calidad del aire ambiente, y para el ozono, la norma actual[2] fijada en 2008 es que la "cuarta concentración máxima diaria anual de 8 horas, promediada durante 3 años" no debe superar las 0,075 partes por millón (ppm). Los detalles exactos de cómo se calcula esto no son importantes para este análisis, pero a grandes rasgos, la concentración media de 8 horas no debe ser muy superior a 0,075 ppm (puede ser mayor debido a la forma en que está redactada la norma).

Echemos un vistazo a las mediciones horarias de ozono.

[2] http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_history.html

```
> summary(ozono$Muestra.Medida)
   Min. 1er Qu.            MedianaMedia3er Qu.
                           Máx. 0.00000 0.02000
0.03200 0.03123 0.04200 0.34900
```

Del resumen se desprende que la concentración horaria máxima es bastante elevada (0,349 ppm) pero que, en general, el grueso de la distribución está muy por debajo de 0,075.

Podemos obtener un poco más de detalle sobre la distribución observando los deciles de los datos.

```
> quantile(ozono$Muestra.Medida, seq(0, 1, 0.1))

                        0%10%20%30%40%50%60%70%
0.000 0.010 0.018 0.023 0.028 0.032 0.036 0.040
      80%90%100%
0.044 0.051 0.349
```

Sabiendo que la norma nacional para el ozono es algo así como 0,075, podemos ver en los datos que

- Los datos son al menos del orden de magnitud correcto (es decir, las unidades son correctas)
- El rango de la distribución es más o menos el que cabría esperar, dada la regulación de los niveles de contaminación ambiental
- Algunos niveles horarios (menos del 10%) están por encima de 0,075, pero esto puede ser razonable teniendo en cuenta la redacción de la norma y el cálculo de la media.

4.8 Hacer una parcela

Hacer un gráfico para visualizar los datos es una buena manera de entender mejor la pregunta y los datos. El trazado puede realizarse en diferentes etapas del análisis de los datos. Para

Por ejemplo, el trazado puede producirse en la fase exploratoria o más tarde en la fase de presentación/comunicación.

Hay dos razones fundamentales para hacer un gráfico de los datos. Son la *creación de expectativas* y la *comprobación de las desviaciones de las expectativas*.

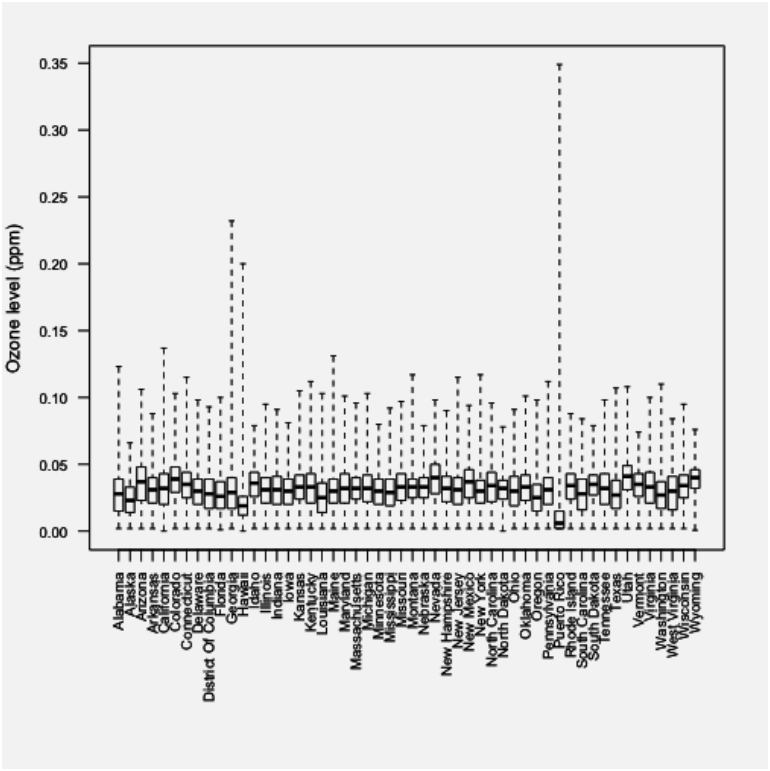
En las primeras etapas del análisis, es posible que disponga de una pregunta/hipótesis, pero que tenga poca idea de lo que ocurre en los datos. Es posible que haya echado un vistazo a algunos de ellos para hacer algunas comprobaciones de cordura, pero si su conjunto de datos es lo suficientemente grande, será difícil simplemente mirar todos los datos. Así que hacer algún tipo de gráfico, que sirva de resumen, será una herramienta útil para *establecer las expectativas de cómo deberían ser los datos*.

Una vez que se han entendido bien los datos, se ha formulado una buena pregunta/hipótesis y se han establecido las expectativas de lo que los datos deberían decir con respecto a la pregunta, hacer un gráfico puede ser una herramienta útil para ver si los datos se ajustan a las expectativas. Los gráficos son especialmente útiles para ver *las desviaciones* de lo que se espera. Las tablas suelen ser buenas para *resumir* los datos presentando cosas como medias, medianas u otras estadísticas. Los gráficos, sin embargo, pueden mostrarte esas cosas, así como mostrarte cosas que están lejos de la media o la mediana, para que puedas comprobar si algo se *supone* que está tan lejos. A menudo, lo que es obvio en un gráfico puede estar oculto en una tabla.

He aquí un simple boxplot[3] de los datos del ozono, con un boxplot para cada estado.

[3]https://en.wikipedia.org/wiki/Box_plot

```
> par(las = 2, mar = c(10, 4, 2, 2), cex.axis = 0.8)
> boxplot(Muestra.Medida ~ Estado.Nombre, ozono, rango = 0, ylab =
"Nivel de ozono (ppm)")
```



Boxplot de los valores de ozono por estado

En el gráfico, podemos ver que para la mayoría de los estados los datos están dentro de un rango bastante estrecho por debajo de 0,05 ppm. Sin embargo, en el caso de Puerto Rico, vemos que los valores típicos son muy bajos, salvo algunos valores extremadamente altos. Del mismo modo, Georgia y Hawái parecen experimentar un valor ocasional muy alto. Puede que merezca la pena investigar más a fondo, en función de su pregunta.

4.9 Pruebe primero la solución fácil

Recordemos que nuestra pregunta original era

> ¿Los condados del este de Estados Unidos tienen niveles de ozono más altos que los del oeste?

¿Cuál es la respuesta más sencilla que podríamos dar a esta pregunta? Por el momento, no te preocupes por si la respuesta es correcta, sino que la cuestión es cómo podrías aportar pruebas *prima facie* de tu hipótesis o pregunta. Más adelante podrá refutar esas pruebas con un análisis más profundo, pero ésta es la primera pasada. Es importante saber que si no se encuentran pruebas de una señal en los datos utilizando un simple gráfico o análisis, entonces es poco probable que se encuentre algo utilizando un análisis más sofisticado.

En primer lugar, tenemos que definir qué entendemos por "este" y "oeste". Lo más sencillo es dividir el país en este y oeste utilizando un valor de longitud específico. Por ahora, utilizaremos -100 como límite. Cualquier monitor con una longitud inferior a -100 será "oeste" y cualquier monitor con una longitud mayor o igual a -100 será "este".

```
> biblioteca(mapas)
> map("estado")
> abline(v = -100, lwd = 3)
> text(-120, 30, "West")
> text(-75, 30, "Este")
```

Mapa de las regiones Este y Oeste

Aquí creamos una nueva variable llamada `región` que utilizamos para indicar si una determinada medición del conjunto de datos se registró en el "este" o en el "oeste".

```
> ozone$region <- factor(ifelse(ozone$Longitude < -100, "west", "eas\
t"))
```

Ahora, podemos hacer un simple resumen de los niveles de ozono en el este y el oeste de los EE.UU. para ver dónde los niveles tienden a ser más altos.

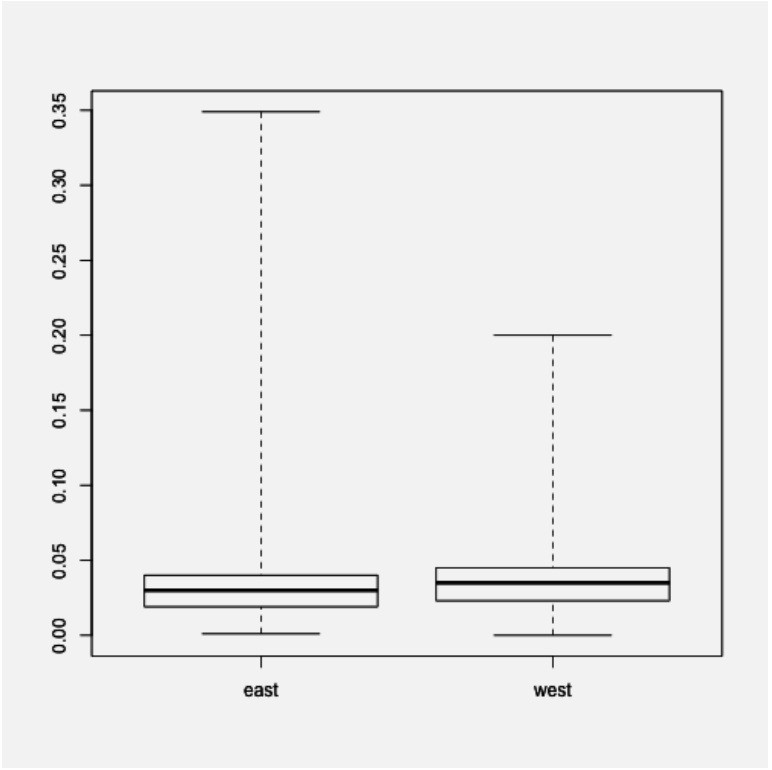
```
> group_by(ozono, región) %>%
+resumir  (media = media(Muestra.Medida, na.rm = TRUE),
+mediana            = mediana(Muestra.Medida, na.rm = TRUE\N-)
))
Fuente: marco de datos local [2 x 3]

               regiónmedia mediana
    (fctr)(dbl)(dbl) 1este
    0,02995250 0,030
2 oeste     0,03400735 0,035
```

Tanto la media como la mediana del nivel de ozono son más altas en el oeste de Estados Unidos que en el este, en aproximadamente 0,004 ppm.

También podemos hacer un boxplot del ozono en las dos regiones para ver cómo se comparan.

```
> boxplot(Muestra.Medida ~ región, ozono, rango = 0)
```



Boxplot del ozono para las regiones este y oeste

Podemos ver en los gráficos de caja que la variabilidad del ozono en el este tiende a ser mucho mayor que la variabilidad en el oeste.

Desafíe su solución

La solución fácil está bien porque es, bueno, fácil, pero nunca hay que permitir que esos resultados se impongan. Siempre hay que pensar en formas de desafiar los resultados, especialmente si esos resultados se ajustan a las expectativas previas.

Recordemos que anteriormente observamos que tres estados tenían algún

valores inusualmente altos de ozono. No sabemos si estos valores son reales o no (por ahora, asumamos que son reales), pero podría ser interesante ver si el mismo patrón de este/oeste se mantiene si eliminamos estos estados que tienen una actividad inusual.

```
> filter(ozono, Estado.Nombre != "Puerto Rico"
         +&State.Name != "Georgia"
         +&State.Name != "Hawaii") %>%
+group_by (region) %>%
+resumir  (media = media(Muestra.Medida, na.rm = TRUE),
+mediana            = mediana(Muestra.Medida, na.rm = TRUE\N-)
))
Fuente: marco de datos local [2 x 3]

               regiónmedia mediana
    (fctr)(dbl)(dbl) 1este
    0,03003692 0,030
2oeste      0,03406880 0,035
```

De hecho, parece que el patrón es el mismo incluso con esos 3 estados eliminados.

4.10 Preguntas de seguimiento

En este capítulo hemos presentado algunos pasos sencillos para empezar un análisis exploratorio. El ejemplo de análisis realizado en este capítulo dista mucho de ser perfecto, pero nos hizo pensar en los datos y en la pregunta de interés. También nos dio una serie de cosas para seguir en caso de que sigamos interesados en esta cuestión.

Llegados a este punto, conviene plantearse algunas preguntas de seguimiento.

1. **¿Tiene los datos adecuados?** A veces, al final de un análisis exploratorio de datos, se llega a la conclusión de que el conjunto de datos no es realmente apropiado para este

pregunta. En este caso, el conjunto de datos parecía perfectamente adecuado para responder a la pregunta de si los condados del este de EE.UU. tienen niveles más altos que los del oeste.

2. **¿Necesita otros datos?** Aunque los datos parecían adecuados para responder a la pregunta planteada, cabe señalar que el conjunto de datos solo abarcaba un año (2014). Puede que merezca la pena examinar si el patrón este/oeste se mantiene para otros años, en cuyo caso tendríamos que salir a buscar otros datos.
3. **¿Tiene la pregunta correcta?** En este caso, no está claro que la pregunta que intentamos responder tenga una relevancia inmediata, y los datos no indicaban realmente nada que aumentara la relevancia de la pregunta. Por ejemplo, podría haber sido más interesante evaluar qué condados infringían la norma nacional de calidad del aire am- biente, porque determinar esto podría tener implicaciones reguladoras. Sin embargo, este es un cálculo mucho más complicado de realizar, ya que requiere datos de al menos 3 años anteriores.

El objetivo del análisis exploratorio de datos es hacernos pensar en los datos y razonar sobre nuestra pregunta. Llegados a este punto, podemos afinar nuestra pregunta o recoger nuevos datos, todo ello en un proceso iterativo para llegar a la verdad.

5. Uso de modelos para explorar los datos

Los objetivos de este capítulo son describir qué es el concepto de modelo de forma más general, explicar cuál es la finalidad de un modelo con respecto a un conjunto de datos y, por último, describir el proceso por el que un analista de datos crea, evalúa y perfecciona un modelo. En un sentido muy general, un modelo es algo que construimos para ayudarnos a entender el mundo real. Un ejemplo común es el uso de un animal que imita una enfermedad humana para ayudarnos a entender, y esperamos, prevenir y/o tratar la enfermedad. El mismo concepto se aplica a un conjunto de datos: es de suponer que se utilizan para comprender el mundo real.

En el mundo de la política, un encuestador tiene un conjunto de datos sobre una muestra de posibles votantes y su trabajo consiste en utilizar esta muestra para predecir el resultado de las elecciones. El analista de datos utiliza los datos del sondeo para construir un modelo que prediga lo que ocurrirá el día de las elecciones. El proceso de construcción de un modelo implica imponer una estructura específica a los datos y crear un resumen de los mismos. En el ejemplo de los datos de las encuestas, puede haber miles de observaciones, por lo que el modelo es una ecuación matemática que refleja la forma o el patrón de los datos, y la ecuación permite resumir las miles de observaciones con, por ejemplo, un número, que podría ser el porcentaje de votantes que votarán a su candidato. Ahora mismo, estos últimos conceptos pueden ser un poco confusos, pero se aclararán mucho más a medida que vayas leyendo.

Un modelo estadístico sirve para dos propósitos clave en un análisis de datos, que son proporcionar un *resumen cuantitativo* de su

datos e imponer una *estructura específica* a la población de la que se tomaron los datos. A veces es útil entender qué es un modelo y por qué puede ser útil mediante la ilustración de ejemplos extremos. El "modelo" trivial **no** es **en absoluto un modelo**.

Imagina que quieres realizar una encuesta a 20 personas para preguntarles cuánto estarían dispuestas a gastar en un producto que estás desarrollando. ¿Cuál es el objetivo de esta encuesta? Previsiblemente, si está invirtiendo tiempo y dinero en el desarrollo de un nuevo producto, cree que hay una gran *población* de personas dispuestas a comprarlo. Sin embargo, es demasiado costoso y complicado preguntar a todos los miembros de esa población lo que estarían dispuestos a pagar. Así que se toma una *muestra* de esa población para tener una idea de lo que pagaría la población.

Uno de nosotros (Roger) ha publicado recientemente un libro titulado *R Pro- gramming for Data Science*[1]. Antes de que se publicara el libro, los lectores interesados podían enviar su nombre y su correo electrónico al sitio web del libro para que se les notificara su publicación. Además, había una opción para especificar cuánto estarían dispuestos a pagar por el libro. A continuación se muestra una muestra de 20 respuestas de personas que ofrecieron esta información.

```
25 20 15 5 30 7 5 10 12 40 30 30 10 25 10 20 10 10 25 5
```

Ahora supongamos que alguien te pregunta: "¿Qué dicen los datos?". Una cosa que podrías hacer es simplemente entregar los datos: los 20 números. Como el conjunto de datos no es tan grande, no sería una gran carga. En última instancia, la respuesta a su pregunta está en ese conjunto de datos, pero tener todos los datos no es un resumen de ningún tipo. Tener todos los datos es importante, pero

[1] https://leanpub.com/rprogramming

no suele ser muy útil. Esto se debe a que el modelo trivial no proporciona ninguna reducción de los datos.

El primer elemento clave de un modelo estadístico es la *reducción de datos*. La idea básica es que se quiere tomar el conjunto original de números que componen el conjunto de datos y transformarlos en un conjunto más pequeño de números. Si originalmente comenzó con 20 números, su modelo debe producir un resumen que tenga menos de 20 números. El proceso de reducción de datos suele terminar con una *estadística*. En general, una estadística es cualquier resumen de los datos. La media de la muestra, o el promedio, es una estadística. También lo son la mediana, la desviación estándar, el máximo, el mínimo y el rango. Algunas estadísticas son más o menos útiles que otras, pero todas son resúmenes de los datos.

Quizás la reducción de datos más sencilla que se puede hacer es la media, o la simple media aritmética, de los datos, que en este caso es de 17,2 dólares. Pasar de 20 números a 1 número es la máxima reducción que se puede hacer en este caso, por lo que definitivamente satisface el elemento de resumen de un modelo.

5.1 Modelos como expectativas

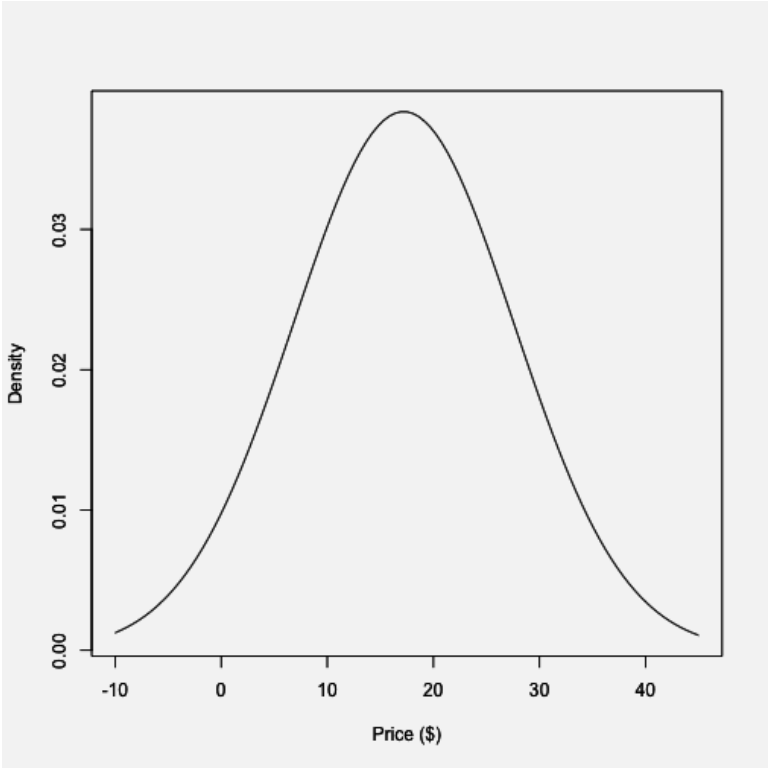
Pero una simple estadística de resumen, como la media de un conjunto de números, no es suficiente para formular un modelo. Un modelo estadístico también debe imponer alguna estructura a los datos. En esencia, **un modelo estadístico proporciona una descripción de cómo funciona el mundo y cómo se generaron los datos**. El modelo es esencialmente una *expectativa* de las relaciones entre varios factores en el mundo real y en su conjunto de datos. Lo que hace que un modelo sea un *modelo estadístico* es que permite cierta aleatoriedad en la generación de los datos.

Aplicación del modelo normal

Quizá el modelo estadístico más popular del mundo sea el modelo normal. Este modelo dice que la aleatoriedad de un conjunto de datos puede explicarse mediante la distribución Normal, o una curva en forma de campana. La distribución Normal está totalmente especificada por dos parámetros: la media y la desviación estándar.

Tomemos los datos que hemos descrito en el apartado anterior: la cantidad de dinero que 20 personas estaban dispuestas a pagar por un hipotético nuevo producto. La esperanza es que estas 20 personas sean una muestra representativa de toda la población que podría comprar este nuevo producto. Si es así, la información contenida en el conjunto de datos puede decir algo sobre todos los miembros de la población.

Para aplicar el modelo normal a este conjunto de datos, sólo tenemos que calcular la media y la desviación estándar. En este caso, la media es de 17,2 dólares y la desviación típica de 10,39 dólares. Teniendo en cuenta estos parámetros, nuestra expectativa según el modelo Normal es que la distribución de los precios que la gente está dispuesta a pagar se parece a esto.



Modelo normal de precios

Según el modelo, cerca del 68% de la población estaría dispuesta a pagar entre 6,81 y 27,59 dólares por este nuevo producto. Que sea una información útil o no depende de los detalles de la situación, que por el momento vamos a pasar por alto.

Si lo desea, puede utilizar el modelo estadístico para responder a preguntas más complejas. Por ejemplo, supongamos que queremos saber "¿Qué proporción de la población estaría dispuesta a pagar más de 30 dólares por este libro?" Usando las propiedades de la distribución Normal (y un poco de ayuda computacional de R), podemos hacer fácilmente este cálculo.

```
pnorm(30, media = media(x), sd = sd(x), lower.tail = FALSE)
```

```
[1] 0.1089893
```

Así que alrededor del 11% de la población estaría dispuesta a pagar más de 30 dólares por el producto. Una vez más, que esto sea útil para ti depende de tus objetivos específicos.

Fíjate en que en la imagen de arriba falta una cosa crucial: ¡los datos! Eso no es exactamente cierto, porque hemos utilizado los datos para hacer el dibujo (para calcular la media y la desviación estándar de la distribución Normal), pero en última instancia los datos no aparecen directamente en el gráfico. En este caso **estamos utilizando la distribución Normal para saber cómo es la población**, no cómo son los datos.

El punto clave aquí es que utilizamos la distribución Normal para establecer la forma de la distribución que *esperamos* que sigan los datos. La distribución Normal es nuestra expectativa de cómo deberían ser los datos.

5.2 Comparación de las expectativas del modelo con la realidad

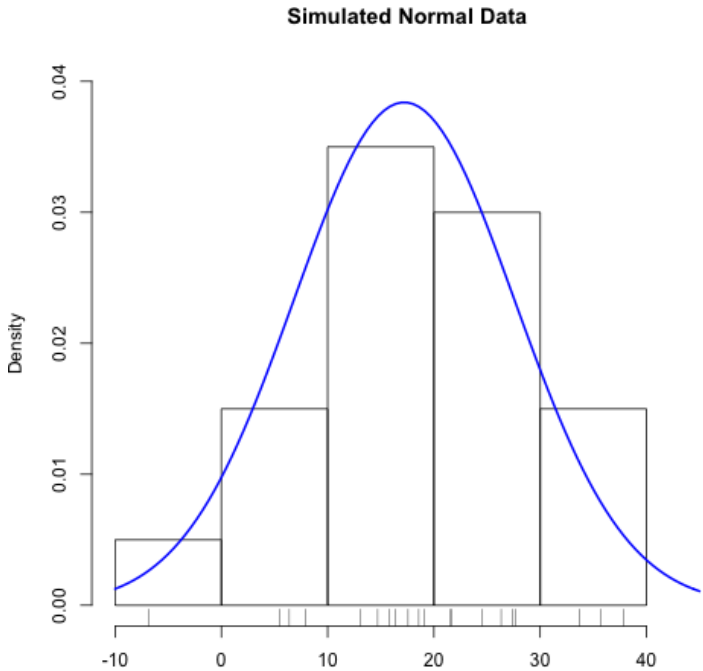
Podemos estar muy orgullosos de haber desarrollado nuestro modelo estadístico, pero en última instancia su utilidad dependerá de lo mucho que refleje los datos que recojamos en el mundo real. ¿Cómo sabemos si nuestras expectativas coinciden con la realidad?

Dibujar una imagen falsa

Para empezar podemos hacer algunas imágenes, como un histograma de los datos. Pero antes de llegar a los datos, vamos a averiguar

lo que *esperamos* ver en los datos. Si la población siguiera aproximadamente una distribución normal y los datos fueran una muestra aleatoria de esa población, entonces la distribución estimada por el histograma debería parecerse al modelo teórico proporcionado por la distribución normal.

En la imagen siguiente, he simulado 20 puntos de datos de una distribución Normal y he superpuesto la curva Normal teórica sobre el histograma.



Histograma de datos normales simulados

Fíjese en lo mucho que coinciden las barras del histograma y la curva azul. Esto es lo que queremos ver con los datos. Si vemos

esto, entonces podríamos concluir que la distribución Normal es un **buen modelo estadístico para los datos**.

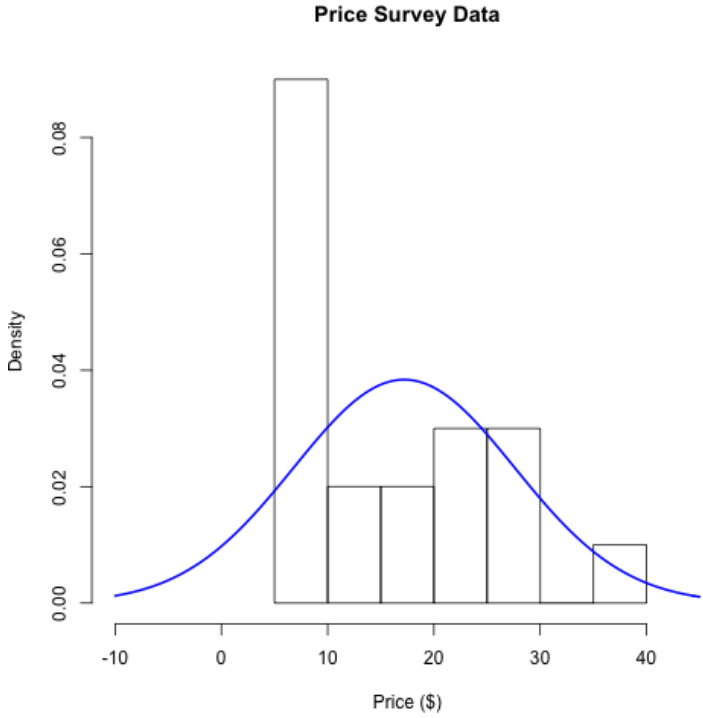
Simular los datos de un modelo hipotético, si es posible, es una buena manera de establecer las expectativas *antes de* ver los datos. Hacer un dibujo falso (incluso a mano, si es necesario) puede ser una herramienta muy útil para iniciar debates sobre el modelo y lo que esperamos de la realidad.

Por ejemplo, antes incluso de examinar los datos, podríamos sospechar que el modelo Normal no ofrece una representación perfecta de la población. En particular, la distribución Normal permite valores *negativos*, pero no esperamos que la gente diga que estaría dispuesta a pagar dólares negativos por un libro.

Así que ya tenemos algunas pruebas de que el modelo normal puede no ser un modelo perfecto, pero ningún modelo es perfecto. La cuestión es si el modelo estadístico proporciona una aproximación razonable que pueda ser útil de alguna manera.

La imagen real

Aquí se muestra un histograma de los datos de la muestra de 20 encuestados. Encima del histograma, he superpuesto la curva normal sobre el histograma de los 20 puntos de datos de la cantidad que la gente dice estar dispuesta a pagar por el libro.



Histograma de los datos de la encuesta de precios

Lo que cabría *esperar* es que el histograma y la línea azul se siguieran más o menos. ¿Cómo se comparan el modelo y la realidad?

A primera vista, parece que el histograma y la distribución normal no coinciden muy bien. El histograma tiene un gran pico alrededor de los 10 dólares, una característica que no está presente en la curva azul. Además, la distribución Normal permite valores negativos en el lado izquierdo del gráfico, pero no hay puntos de datos en esa región del gráfico.

Hasta ahora los datos sugieren que el modelo Normal no es realmente una muy buena representación de la población, dados los datos

que hemos tomado como muestra de la población. Parece que las 20 personas encuestadas tienen una fuerte preferencia por pagar un precio en torno a los 10 dólares, mientras que hay unas pocas personas dispuestas a pagar más que eso. Estos rasgos de los datos no están bien caracterizados por una distribución normal.

5.3 Reaccionar ante los datos: Afinando nuestras expectativas

Bien, el modelo y los datos no coinciden muy bien, como indica el histograma anterior. Entonces, ¿qué hacer? Bueno, podemos

1. Conseguir un modelo diferente; o
2. Obtener datos diferentes

O podríamos hacer ambas cosas. Lo que hagamos en respuesta depende un poco de nuestras creencias sobre el modelo y de nuestra comprensión del proceso de recogida de datos. Si creemos firmemente que la población de precios que la gente estaría dispuesta a pagar debería seguir una distribución Normal, entonces sería menos probable que hiciéramos modificaciones importantes en el modelo. Podríamos examinar el proceso de recogida de datos para ver si tal vez ha provocado algún sesgo en los datos. Sin embargo, si el proceso de recogida de datos es correcto, podríamos vernos obligados a reexaminar nuestro modelo para la población y ver qué se podría cambiar. En este caso, es probable que nuestro modelo sea inadecuado, sobre todo teniendo en cuenta que es difícil imaginar un proceso de recogida de datos válido que pueda dar lugar a valores negativos en los datos (como permite la distribución Normal).

Para cerrar el círculo aquí, elegiremos un modelo estadístico diferente para representar a la población, la

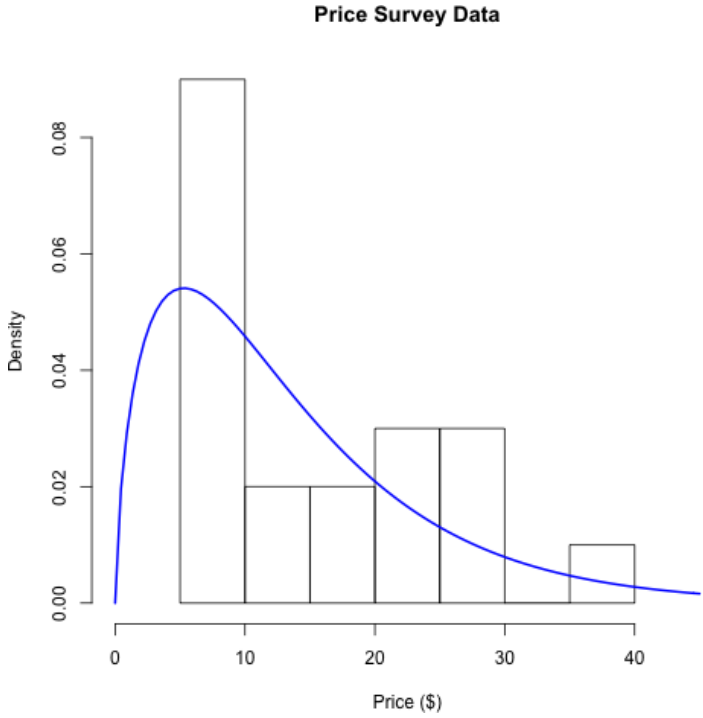
distribución Gamma. Esta distribución tiene la característica de que sólo admite valores positivos

por lo que se elimina el problema que teníamos con los valores negativos con la distribución Normal.

Ahora, debemos volver a la parte superior de nuestra iteración y hacer lo siguiente:

1. Desarrollar las expectativas: Dibuja una imagen falsa: ¿qué esperamos ver antes de mirar los datos?
2. Comparar nuestras expectativas con los datos
3. Afinar nuestras expectativas, teniendo en cuenta lo que muestran los datos

Para su referencia, aquí hay un histgrama de los mismos datos con la distribución Gamma (estimada usando los datos) superpuesta.



Datos de la encuesta de precios con distribución gamma

¿Cómo se ajustan ahora los datos a sus expectativas?

Se preguntará qué diferencia hay entre el modelo que utilizo para representar a la población de la que proceden los datos. Bueno, para empezar, puede afectar al tipo de predicciones que se pueden hacer con el modelo. Por ejemplo, recordemos que antes nos interesaba saber qué proporción de la población estaría dispuesta a pagar al menos 30 dólares por el libro. Nuestro nuevo modelo dice que sólo un 7% de la población estaría dispuesta a pagar al menos esa cantidad (el modelo normal afirmaba que el 11% pagaría 30 dólares o más). Por tanto, los distintos modelos pueden dar lugar a predicciones diferentes en función de

los mismos datos, lo que puede repercutir en las decisiones que se tomen en el futuro.

5.4 Examinar las relaciones lineales

Es habitual observar los datos y tratar de entender las relaciones lineales entre las variables de interés. La técnica estadística más común para ayudar en esta tarea es la *regresión lineal*. Podemos aplicar los principios que hemos discutido anteriormente - desarrollar expectativas, comparar nuestras expectativas con los datos, refinar nuestras expectativas - también a la aplicación de la regresión lineal.

Para este ejemplo veremos un simple conjunto de datos de calidad del aire que contiene información sobre los niveles de ozono troposférico en la ciudad de Nueva York en el año 1999 para los meses de mayo a 1999. Aquí están las primeras filas del conjunto de datos.

```
          ozonetemp   mes
1 25.37262 55.33333     5
2 32.83333 57.66667     5
3 28.88667 56.66667     5
4 12.06854 56.66667     5
5 11.21920 63.66667     5
6 13.19110 60.00000     5
```

Los datos contienen los niveles medios diarios de ozono (en partes por billón [pbb]) y la temperatura (en grados Fahrenheit). Una pregunta de interés que podría motivar la recopilación de este conjunto de datos es "¿Cómo se relaciona la temperatura ambiental con los niveles de ozono en Nueva York?"

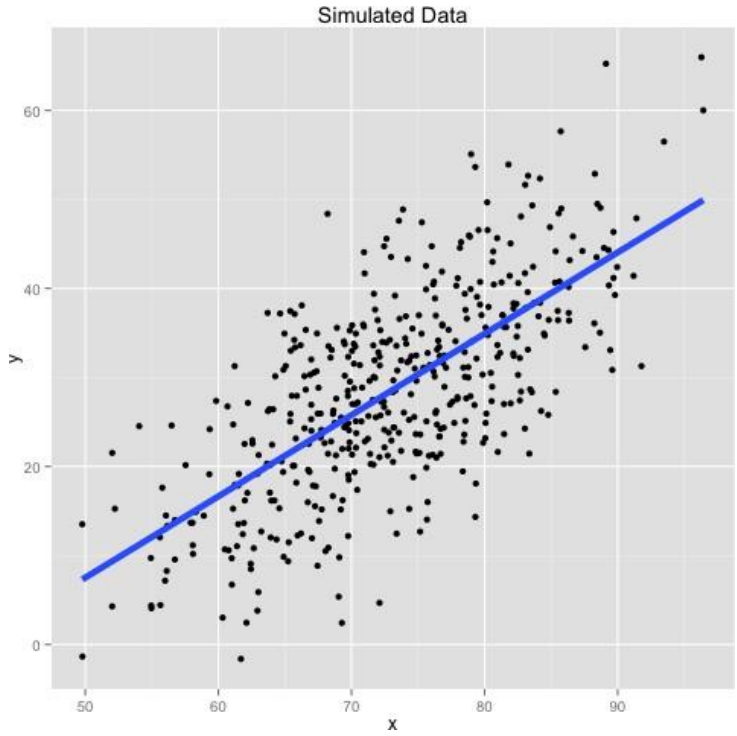
Expectativas

Después de leer un poco sobre la formación de ozono en la atmósfera[2]sabemos que la formación de ozono depende fundamentalmente de la presencia de luz solar. La luz solar también está relacionada con la temperatura en el sentido de que en los días en los que hay mucha luz solar, es de esperar que la temperatura media de ese día sea más alta. Los días nublados tienen una temperatura media más baja y menos ozono. Por lo tanto, hay razones para creer que en los días con temperaturas más altas se espera que haya mayores niveles de ozono. Se trata de una relación indirecta, ya que utilizamos la temperatura como indicador de la cantidad de luz solar.

El modelo más sencillo que podemos formular para caracterizar la relación entre la temperatura y el ozono es un *modelo lineal*. Este modelo dice que a medida que aumenta la temperatura, la cantidad de ozono en la atmósfera aumenta linealmente con ella. ¿Cómo esperamos que sea esto?

Podemos simular algunos datos para hacer una *imagen falsa* de cómo debería ser la relación entre el ozono y la temperatura según un modelo lineal. Aquí tenemos una relación lineal simple junto con los datos simulados en un gráfico de dispersión.

[2]https://en.wikipedia.org/wiki/Tropospheric_ozone



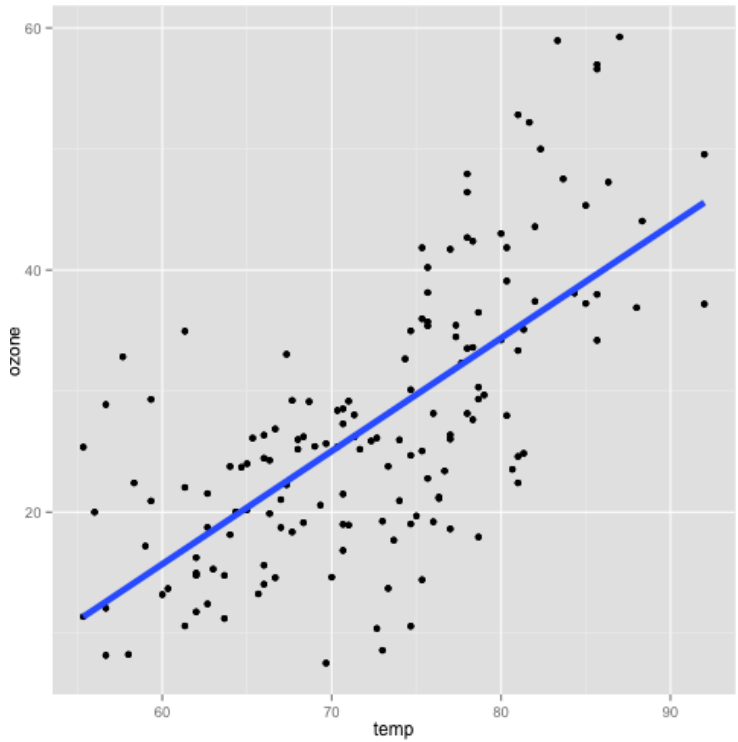
Datos simulados con un modelo lineal

Observe que si elige cualquier punto de la línea azul, hay aproximadamente el mismo número de puntos por encima de la línea que por debajo de ella (esto también se conoce como errores insesgados). Además, los puntos del gráfico de dispersión parecen aumentar linealmente a medida que se avanza hacia la derecha en el eje x, aunque haya bastante ruido/dispersión a lo largo de la línea.

Si estamos en lo cierto con nuestro modelo lineal, y ese es el modelo que caracteriza los datos y la relación entre el ozono y la temperatura, entonces, a grandes rasgos, esta es la imagen que deberíamos ver cuando graficamos los datos.

Comparación de las expectativas con los datos

Esta es la imagen de los datos reales de ozono y temperatura en la ciudad de Nueva York para el año 1999. Sobre el gráfico de dispersión de los datos, hemos representado la línea de regresión lineal ajustada estimada a partir de los datos.



Modelo lineal de ozono y temperatura

¿Cómo se compara esta imagen con la que esperabas ver?

Una cosa está clara: parece que hay una tendencia al aumento del ozono a medida que aumenta la temperatura, tal y como se hipotetiza.

de tamaño. Sin embargo, hay algunas desviaciones con respecto a la bonita imagen falsa que hicimos anteriormente. Los puntos no parecen estar uniformemente equilibrados alrededor de la línea de regresión azul.

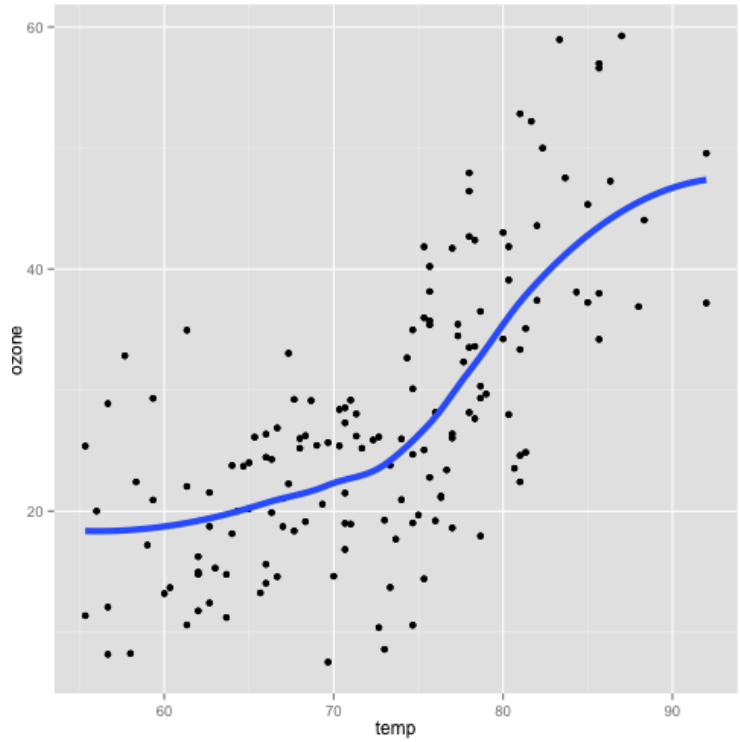
Si se dibuja una línea vertical alrededor de una temperatura de 85 grados, se observa que la mayoría de los puntos están por encima de la línea. Si se dibuja una línea vertical alrededor de 70 grados, se observa que la mayoría de los puntos están por debajo de la línea. Esto implica que a temperaturas más altas, nuestro modelo está sesgado hacia abajo (subestima el ozono) y a temperaturas moderadas nuestro modelo está sesgado hacia arriba. Esto no es una gran característica, en esta situación podríamos preferir que nuestro modelo no esté sesgado en ningún lugar.

Nuestro modelo de regresión lineal simple parece capturar la relación general creciente entre la temperatura y el ozono, pero parece estar sesgado en ciertos rangos de temperatura. Parece que hay margen de mejora con este modelo si queremos caracterizar mejor la relación entre la temperatura y el ozono en este conjunto de datos.

Afinar las expectativas

De la imagen anterior se desprende que la relación entre la temperatura y el ozono puede no ser lineal. De hecho, los puntos de datos sugieren que tal vez la relación sea plana hasta unos 70 grados y que después los niveles de ozono aumentan rápidamente con la temperatura. Esto sugiere una relación *no lineal* entre la temperatura y el ozono.

La forma más fácil de capturar esta expectativa revisada es con un suavizador, en este caso un suavizador de loess.



Alisador de Loess para el ozono y la temperatura

Este gráfico muestra una imagen diferente: la relación aumenta lentamente hasta los 75 grados aproximadamente, y luego aumenta bruscamente. Alrededor de los 90 grados, se sugiere que la relación se nivela de nuevo.

Los suavizadores (como el loess) son herramientas útiles porque captan rápidamente las tendencias de un conjunto de datos sin hacer ninguna suposición estructural sobre los mismos. Básicamente, son una forma automática o informatizada de trazar una curva sobre unos datos. Sin embargo, los suavizadores rara vez dicen algo sobre el mecanismo de la relación, por lo que pueden ser limitados en ese sentido. Para saber más sobre la relación

entre la temperatura y el ozono, puede que tengamos que recurrir a un modelo más detallado que el simple modelo lineal que teníamos antes.

5.5 ¿Cuándo nos detenemos?

En los ejemplos anteriores, hemos completado una iteración del proceso de análisis de datos. En algunos casos, una sola iteración puede ser suficiente, pero en la mayoría de los casos de la vida real, tendrá que iterar al menos unas cuantas veces. De los ejemplos anteriores, todavía quedan algunas cosas por hacer:

- **Datos de la encuesta de precios**: Terminamos el ejemplo ajustando un modelo de distribución Gamma. Pero, ¿cómo se ajusta eso a los datos? ¿Qué esperaríamos de los datos si realmente siguieran una distribución Gamma (nunca hicimos ese gráfico)? ¿Existe una forma mejor de capturar ese pico en la distribución justo alrededor de los 10 dólares?
- **Ozono y temperatura**: El suavizador sugirió una relación no lineal entre la temperatura y el ozono, pero ¿cuál es la razón de ello? ¿Es la no linealidad real o sólo una casualidad en los datos? ¿Existe algún proceso físico conocido que explique la dramática aumento de los niveles de ozono a partir de una determinada temperatura y podemos modelar ese proceso?

En última instancia, se puede iterar una y otra vez. Cada respuesta suele suscitar más preguntas y exige seguir indagando en los datos. Entonces, ¿cuándo se detiene exactamente el proceso? La teoría estadística sugiere una serie de enfoques diferentes para determinar cuándo un modelo estadístico es "suficientemente bueno" y se ajusta bien a los datos. Esto no es lo que discutiremos aquí, sino que discutiremos algunos criterios de alto nivel para determinar cuándo se puede considerar detener la iteración del análisis de datos.

¿Te has quedado sin datos?

El análisis iterativo de los datos acabará planteando preguntas que no pueden responderse con los datos disponibles. Por ejemplo, en el análisis de ozono/temperatura, la modelización sugirió que no hay una relación simple entre las dos variables, que puede ser no lineal. Pero los datos no pueden explicar con precisión por qué puede existir esa relación no lineal (aunque pueden sugerir ciertas hipótesis). Además, es posible que tenga que recoger datos adicionales para determinar si lo que observa es real o simplemente una casualidad o un accidente estadístico. En cualquier caso, hay que volver a salir al mundo y recoger nuevos datos. Es poco probable que más análisis de datos aporten estas respuestas.

Otra situación en la que se puede encontrar buscando más datos es cuando se ha completado el análisis de datos y se ha llegado a resultados satisfactorios, normalmente algún hallazgo interesante. Entonces, puede ser muy importante intentar *replicar* lo que se ha encontrado utilizando un conjunto de datos diferente, posiblemente independiente. En el ejemplo del ozono/temperatura, si llegamos a la conclusión de que existe una relación no lineal entre la temperatura y el ozono, nuestra conclusión podría ser más convincente si pudiéramos demostrar que esta relación está presente en otras ciudades además de Nueva York. Esta confirmación independiente puede aumentar la fuerza de las pruebas y desempeñar un papel importante en la toma de decisiones.

¿Tiene suficientes pruebas para tomar una decisión?

El análisis de datos se lleva a cabo a menudo para apoyar la toma de decisiones, ya sea en los negocios, en el mundo académico, en el gobierno o en cualquier otro lugar, a menudo recogemos y analizamos datos para informar de algún tipo de decisión. Es importante darse cuenta de que el análisis que se realiza para llegar al punto en el que

puede tomar una decisión sobre algo puede ser muy diferente del análisis que se realiza para conseguir otros objetivos, como escribir un informe, publicar un artículo o sacar un producto acabado.

Por eso es importante tener siempre presente el *propósito del análisis de datos a medida que se avanza,* porque se pueden invertir recursos en exceso o en defecto en el análisis si éste no está en sintonía con el objetivo final. El propósito de un análisis de datos puede cambiar con el tiempo y, de hecho, puede haber múltiples propósitos paralelos. La cuestión de si tiene suficientes pruebas depende de factores específicos de la aplicación en cuestión y de su situación personal con respecto a los costes y beneficios. Si cree que no tiene suficientes pruebas para tomar una decisión, puede ser porque no tiene datos o porque necesita realizar más análisis.

¿Puede situar sus resultados en un contexto más amplio?

Otra forma de plantear esta pregunta es: "¿Los resultados tienen algún tipo de sentido?". A menudo, puede responder a esta pregunta buscando en la literatura disponible en su área o ver si otras personas dentro o fuera de su organización han llegado a una conclusión similar. Si los resultados de su análisis se ajustan a lo que otros han encontrado, eso puede ser algo bueno, pero no es el único resultado deseable. Los hallazgos que no coinciden con los resultados anteriores pueden conducir a un nuevo descubrimiento. En cualquier caso, a menudo es difícil llegar a la respuesta correcta sin una investigación más profunda.

Hay que tener un poco de cuidado con la respuesta a esta pregunta. A menudo, especialmente con conjuntos de datos muy grandes y complejos, es fácil llegar a un resultado que "tiene sentido" y se ajusta a nuestra comprensión de cómo *debería* funcionar un proceso determinado. En esta situación, es importante ser hipercrítico con nuestros resultados y cuestionarlos en la medida de lo posible. En nuestro ex

periencia, cuando los datos se ajustan mucho a nuestras expectativas, puede ser el resultado de errores o malentendidos en el análisis o en el proceso de recogida de datos. Es fundamental cuestionar todos los aspectos del proceso de análisis para asegurarse de que todo se ha hecho correctamente.

Si los resultados *no* tienen sentido, o los datos no coinciden con sus expectativas, aquí es donde las cosas se ponen interesantes. Puede que simplemente hayas hecho algo incorrecto en el análisis o en la recogida de datos. Lo más probable es que eso sea exactamente lo que ha ocurrido. Por cada diamante en bruto, hay 99 trozos de carbón. Sin embargo, en el caso de que haya descubierto algo inusual que otros no hayan visto todavía, tendrá que (a) asegurarse de que el análisis se hizo correctamente y (b) replicar sus hallazgos en otro conjunto de datos. Los resultados sorprendentes suelen ser objeto de mucho escrutinio y tendrás que estar preparado para defender rigurosamente tu trabajo.

En última instancia, si su análisis le lleva a un lugar en el que puede responder definitivamente a la pregunta "¿Tienen sentido los resultados?", entonces, independientemente de cómo responda a esa pregunta, es probable que tenga que **detener su análisis y comprobar cuidadosamente cada parte del mismo**.

¿Se te ha acabado el tiempo?

Este criterio parece arbitrario pero, sin embargo, desempeña un papel importante a la hora de determinar cuándo hay que detener un análisis en la práctica. Una pregunta relacionada podría ser "¿Se ha quedado sin dinero?". En última instancia, habrá un presupuesto de tiempo y un presupuesto monetario que determinará cuántos recursos se pueden dedicar a un análisis determinado. Estar al tanto de estos presupuestos, aunque no los controle necesariamente, puede ser importante para gestionar un análisis de datos. En particular, es posible que tenga que argumentar para obtener más recursos y

persuadir a otros para que se los den. En una situación así,

es útil saber cuándo hay que detener la iteración del análisis de datos y preparar los resultados que se hayan obtenido hasta la fecha para presentar un argumento coherente para la continuación del análisis.

5.6 Resumen

La creación de modelos, como todo el proceso de análisis de datos, es un proceso iterativo. Los modelos se utilizan para proporcionar una reducción de los datos y para dar una idea de la población sobre la que se intenta hacer una inferencia. Es importante establecer primero sus expectativas sobre cómo un modelo debe caracterizar un conjunto de datos antes de aplicar realmente un modelo a los datos. A continuación, puede comprobar si su modelo se ajusta a sus expectativas. A menudo, habrá características del conjunto de datos que no se ajusten a su modelo y tendrá que refinar su modelo o examinar el proceso de recogida de datos.

6. Inferencia: Un manual

La inferencia es uno de los muchos objetivos posibles en el análisis de datos, por lo que vale la pena discutir qué es exactamente el acto de hacer una inferencia. Recordemos que anteriormente describimos que uno de los seis tipos de preguntas que se pueden hacer en un análisis de datos es una pregunta **inferencial**. Entonces, ¿qué es la inferencia?

En general, el objetivo de la inferencia es poder hacer una afirmación sobre algo que *no se ha observado* y, en el mejor de los casos, poder caracterizar cualquier incertidumbre que se tenga sobre esa afirmación. La inferencia es difícil debido a la diferencia entre lo que se puede observar y lo que en última instancia se quiere saber.

6.1 Identificar la población

El lenguaje de la inferencia puede cambiar según la aplicación, pero lo más habitual es que nos refiramos a las cosas que no podemos observar (pero de las que queremos saber) como la **población** o como características de la población y a los datos que observamos como la **muestra**. El objetivo es utilizar la muestra para hacer alguna afirmación sobre la población. Para ello, tenemos que especificar algunas cosas.

La identificación de la población es la tarea más importante. Si no se puede identificar o describir coherentemente la población, no se puede hacer una inferencia. Basta con detenerse. Una vez que hayas averiguado cuál es la población y sobre qué característica de la población quieres hacer una afirmación (por ejemplo, la media), entonces podrás traducirlo más tarde en una

utilizando un modelo estadístico formal (que se trata más adelante en este libro).

6.2 Describir el proceso de muestreo

¿Cómo llegaron los datos desde la población a su ordenador? Poder describir este proceso es importante para determinar si los datos son útiles para hacer inferencias sobre las características de la población. Como ejemplo extremo, si está interesado en la edad media de las mujeres en una población, pero su proceso de muestreo está diseñado de alguna manera para que sólo produzca datos sobre los hombres, entonces no puede utilizar los datos para hacer una inferencia sobre la edad media de las mujeres. Entender el proceso de muestreo es clave para determinar si la muestra es *representativa* de la población de interés. Tenga en cuenta que si tiene dificultades para describir la población, tendrá dificultades para describir el proceso de muestreo de datos de la población. Por lo tanto, la descripción del proceso de muestreo depende de su capacidad para describir coherentemente la población.

6.3 Describir un modelo para la población

Necesitamos tener una representación abstracta de cómo se relacionan los elementos de la población entre sí. Por lo general, se trata de un modelo estadístico que podemos representar mediante notación matemática. Sin embargo, en situaciones más complejas, podemos recurrir a representaciones algorítmicas que no pueden escribirse claramente en papel (muchos enfoques de aprendizaje automático tienen que describirse de esta manera).

El modelo más sencillo podría ser un *modelo lineal simple*, como

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Aquí, x e y son características de la población y β_0 y β_1 describen la relación entre esas características (es decir, ¿están asociadas positiva o negativamente?). El elemento final ε es un cajón de sastre que pretende capturar todos los factores que contribuyen a la diferencia entre y y lo que *esperamos* que sea y, que es $\beta_0 + \beta_1 x$. Es esta última parte la que hace que el modelo sea un modelo estadístico, porque normalmente permitimos que ε sea aleatorio.

Otra característica sobre la que solemos tener que hacer una suposición es cómo interactúan las diferentes unidades de la población entre sí. Normalmente, sin ninguna información adicional, supondremos que las unidades de la población son *independientes*, lo que significa que las mediciones de una unidad no proporcionan ninguna información sobre las mediciones de otra unidad. En el mejor de los casos, esta suposición es aproximadamente cierta, pero puede ser una aproximación útil. En algunas situaciones, como cuando se estudian cosas que están estrechamente conectadas en el espacio o el tiempo, la suposición es claramente falsa, y debemos recurrir a enfoques especiales de modelización para tener en cuenta la falta de independencia.

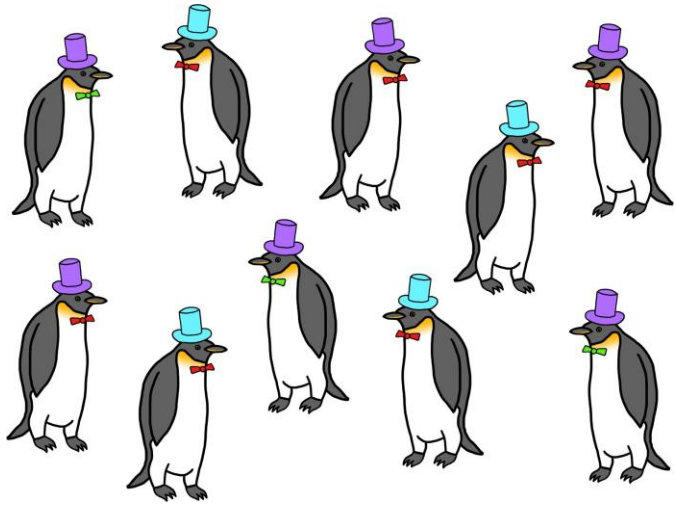
George Box, un estadístico, dijo una vez que[1] "todos los modelos son erróneos, pero algunos son útiles". Es probable que, sea cual sea el modelo que conciba para describir las características de una población, sea técnicamente incorrecto. Pero no hay que obsesionarse con el desarrollo de un modelo *correcto*, sino que hay que identificar un modelo que sea útil y que cuente una historia sobre los datos y sobre los procesos subyacentes que se intentan estudiar.

6.4 Un ejemplo rápido

Piensa en este grupo de pingüinos (porque los pingüinos son increíbles), cada uno de ellos con un sombrero púrpura o turquesa.

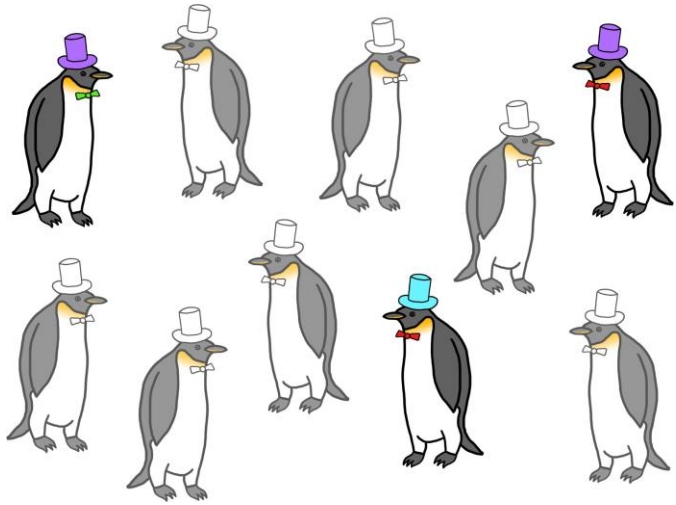
[1] https://en.wikipedia.org/wiki/All_models_are_wrong

Hay un total de 10 pingüinos en este grupo. Los llamaremos la *población*.



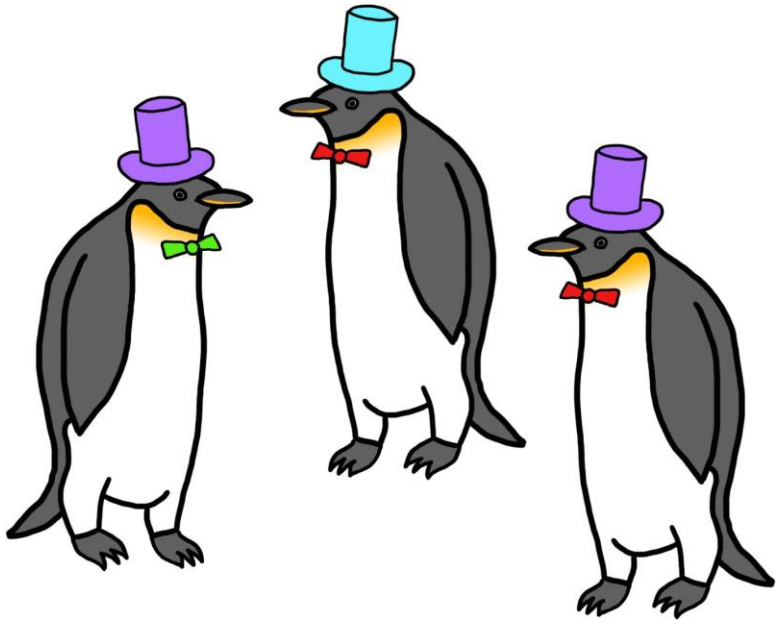
Población de pingüinos con sombreros turquesa y púrpura

Ahora supongamos que quieres saber qué proporción de la *población* de pingüinos lleva sombreros turquesa. Pero hay una trampa: no tienes tiempo, dinero o capacidad para cuidar de 10 pingüinos. ¿Quién lo tiene? Sólo puedes permitirte cuidar de tres pingüinos, así que tomas una muestra al azar de tres de estos 10 pingüinos.



Muestra de 3 pingüinos de la población

El punto clave es que nunca se observa la población completa de pingüinos. Ahora lo que tienes es tu *conjunto de datos*, que contiene sólo tres pingüinos.



Conjunto de datos de pingüinos

Llegados a este punto, una pregunta fácil de hacer es "¿Qué proporción de los pingüinos *de mi conjunto de datos llevan sombreros turquesa*?". En la imagen anterior, está claro que 1/3 de los pingüinos llevan sombreros turquesa. *No* tenemos ninguna *duda* sobre esa proporción porque los datos están delante de nosotros.

La pregunta difícil es: "Según los datos que tengo, ¿qué proporción de los pingüinos de la *población* original llevan sombreros turquesa?" En este momento, sólo tenemos nuestra muestra de tres pingüinos y no observamos a toda la población. ¿Qué podemos hacer? Tenemos que hacer una *inferencia* sobre la población utilizando los datos que tenemos a mano.

Las tres cosas que tenemos que hacer para hacer una inferencia son:

1. **Definir la población**. En este caso, la población son los 10 pingüinos originales de los que tomamos una muestra de nuestro conjunto de datos de tres pingüinos.
2. **Describe el proceso de muestreo**. No lo hemos mencionado explícitamente, pero supongamos por ahora que nuestro "proceso de muestreo" consistió en tomar los tres primeros pingüinos que se acercaron a nosotros.
3. **Describe un modelo para la población**. Supondremos que los sombreros que llevan los pingüinos son *independientes* entre sí, de modo que el hecho de que un pingüino tenga un sombrero morado no influye en que otro tenga un sombrero turquesa. Como sólo queremos estimar una proporción simple de pingüinos con sombreros turquesa, no necesitamos hacer ninguna suposición más compleja sobre cómo se relacionan los pingüinos entre sí.

Dados los tres ingredientes anteriores, podríamos estimar que la proporción de pingüinos con sombreros turquesa es de 1/3. ¿Qué tan buena es esta estimación? Dado que conocemos la verdad
aquí-2/5 de los pingüinos tienen sombreros turquesa en la población- podríamos preguntar si 1/3 es una estimación razonable o no.
La respuesta a esta pregunta depende de una serie de factores que se analizarán en la siguiente sección.

6.5 Factores que afectan a la calidad de la inferencia

Los factores clave que afectan a la calidad de una inferencia que se pueda hacer están relacionados con las violaciones de nuestro pensamiento sobre el proceso de muestreo y el modelo de la población. Obviamente, si no podemos definir la población de forma coherente,

cualquier "inferencia" que hagamos sobre la población estará igualmente definida de forma imprecisa.

Si no entendemos cómo funciona el proceso de muestreo, los datos que recojamos no representarán a la población como pensábamos. Esto afectaría a nuestra inferencia, ya que la inferencia que haríamos no se aplicaría a toda la población, sino a una selección específica de la misma. Este fenómeno se denomina a veces **sesgo de selección**, ya que las cantidades que se estiman están sesgadas hacia la selección de la población que se *ha* muestreado.

Una violación del modelo que planteamos para la población podría hacer que estimáramos una relación errónea entre las características de la población o que subestimáramos la inseguridad de nuestras estimaciones. Por ejemplo, si es cierto que los pingüinos pueden influir en el color de los sombreros que llevan otros pingüinos, esto violaría el supuesto de independencia entre pingüinos. Esto daría lugar a un aumento de la incertidumbre de cualquier estimación que hagamos a partir de los datos. En general, la dependencia entre las unidades de una población reduce el "tamaño efectivo de la muestra" del conjunto de datos porque las unidades que se observan no son realmente independientes entre sí y no representan fragmentos independientes de información.

Una última razón para la diferencia entre nuestra estimación a partir de los datos y la verdad en la población es la **variabilidad del muestreo**. Dado que muestreamos pingüinos de la población de forma aleatoria, es probable que si volviéramos a realizar el experimento y tomáramos muestras de otros tres pingüinos, obtendríamos una estimación diferente del número de pingüinos con sombreros turquesa, simplemente debido a la variación aleatoria en el proceso de muestreo. Esto ocurriría incluso si nuestra descripción del proceso de muestreo fuera precisa y nuestro modelo para la población fuera perfecto.

En la mayoría de los casos, las diferencias entre lo que podemos estimar con los datos y lo que es la verdad en la población pueden ser

se explica por una combinación de los tres factores. A veces puede ser difícil determinar el papel que desempeña cada uno de ellos en un problema determinado debido a la falta de información, pero suele merecer la pena reflexionar sobre cada uno de estos factores y decidir cuál puede estar desempeñando un papel dominante. De este modo, se podrá corregir el problema, por ejemplo, en futuros estudios o experimentos.

6.6 Ejemplo: Uso de Apple Music

El 18 de agosto de 2015, la empresa de investigación de mercados de consumo MusicWatch publicó un estudio[2] sobre un nuevo servicio de música lanzado por Apple, Inc. llamado Apple Music. El servicio era un nuevo servicio de música en streaming diseñado para dar a los usuarios acceso a un amplio catálogo de música por 9,99 dólares al mes. Sin embargo, había un periodo de prueba gratuito que duraba 3 meses. En aquel momento se especuló mucho sobre cuántos usuarios seguirían pagando los 9,99 dólares al mes una vez terminada la prueba gratuita.

El estudio de MusicWatch afirmaba, entre otras cosas, que

> Entre las personas que han probado Apple Music, el 48 por ciento ha declarado que no utiliza actualmente el servicio.

Esto sugeriría que casi la mitad de las personas que se habían inscrito en el periodo de prueba gratuito de Apple Music no estaban interesadas en seguir utilizándolo y probablemente no pagarían por él una vez finalizado el periodo de prueba. De ser cierto, sería un golpe para el servicio recién lanzado.

[2] http://www.businesswire.com/news/home/20150818005755/en#.VddbR7Scy6F

¿Pero cómo ha llegado MusicWatch a esta cifra? Afirmó haber encuestado a 5.000 personas en su estudio. Poco antes de que se publicara la encuesta de MusicWatch, Apple afirmó que unos 11 millones de personas se habían inscrito en su nuevo servicio Apple Music (como el servicio acababa de lanzarse, todos los que se habían inscrito estaban en el periodo de prueba gratuito). Evidentemente, 5.000 personas no constituyen toda la población, por lo que no tenemos más que una pequeña muestra de usuarios.

¿Cuál es el objetivo al que quería responder MusicWatch? Parece que querían saber el porcentaje de *todas las personas que se habían suscrito a Apple Music* que seguían utilizando el servicio. Como habría sido enormemente caro encuestar a los 11 millones de personas, tuvieron que recurrir a una muestra mucho más pequeña de 5.000. ¿Pueden hacer una inferencia sobre toda la población a partir de la muestra de 5.000?

Consideremos los tres ingredientes de la inferencia:

1. **Población**: Nos interesa el comportamiento de toda la base de usuarios de Apple Music, que son aproximadamente 11 millones de personas, según Apple.
2. **Proceso de muestreo**: El comunicado de prensa no aclara cómo se llevó a cabo el estudio y cómo se recogieron los datos. Es probable que se tratara de una encuesta telefónica, por lo que se seleccionó a personas al azar para llamarlas y preguntarles sobre su uso del servicio. ¿Crees que este proceso dio lugar a una muestra de encuestados que es representativa de toda la población de usuarios de Apple Music?
3. **Modelo para la población**: Dado el tamaño relativamente pequeño de la muestra en relación con toda la población, es probable que se pueda pensar que los individuos de la encuesta son independientes entre sí. En otras palabras, es poco probable que un encuestado de la encuesta pueda haber influido en

otro encuestado.

Si la muestra es representativa y los individuos son independientes, podríamos utilizar la cifra del 48% como estimación del porcentaje de la población que ya no utiliza el servicio. El comunicado de prensa de MusicWatch no indica ninguna medida de incertidumbre, por lo que no sabemos la fiabilidad de la cifra.

Curiosamente, poco después de que se volviera a publicar la encuesta de MusicWatch, Apple hizo una declaración a la publicación *The Verge*, en la que afirmaba que el 79% de los usuarios que se habían inscrito seguían utilizando el servicio (es decir, que sólo el 21% había dejado de utilizarlo, frente al 48% informado por MusicWatch). Ahora bien, la diferencia entre Apple y MusicWatch es que Apple tiene fácil acceso a toda la población de usuarios de Apple Music. Si quieren saber qué porcentaje de la *población* de usuarios sigue utilizándolo, sólo tienen que contar el número de usuarios activos del servicio y dividirlo por el número total de personas que se inscribieron. *No* hay *incertidumbre* sobre esa cifra concreta, porque no se ha necesitado un muestreo para estimarla (supongo que Apple no ha utilizado el muestreo para estimar el porcentaje).

Si creemos que Apple y MusicWatch estaban midiendo lo mismo en sus análisis (y no está claro que lo hicieran), esto sugeriría que la estimación de MusicWatch del porcentaje de población (48%) estaba bastante alejada del valor real (21%). ¿Qué explicaría esta gran diferencia?

1. **Variación aleatoria**. Es cierto que la encuesta de MusicWatch era una muestra pequeña en relación con toda la población, pero la muestra seguía siendo grande, con 5.000 personas. Además, el análisis era bastante sencillo (sólo se tomaba la proporción de usuarios que seguían utilizando el servicio), por lo que es poco probable que la incertidumbre asociada a esa estimación sea tan grande.

2. **Sesgo de selección**. Recordemos que no está claro cómo Mu- sicWatch muestreó a sus encuestados, pero es posible que la forma en que lo hicieron les llevara a capturar un conjunto de encuestados que estaban menos inclinados a usar Apple Music. Más allá de esto, no podemos decir más sin conocer los detalles del proceso de la encuesta.
3. **Diferencias de medición**. Una cosa que no sabemos es cómo definen MusicWatch o Apple "seguir usando el servicio". Se pueden imaginar varias formas de determinar si una persona sigue utilizando el servicio. Se podría preguntar "¿Lo has usado en la última semana?" o quizás "¿Lo usaste ayer?". Las respuestas a estas preguntas serían muy diferentes y probablemente conducirían a diferentes porcentajes globales de uso.
4. **Los encuestados no son independientes**. Es posible que los encuestados no sean independientes entre sí. Esto afectaría principalmente a la incertidumbre sobre la estimación, haciéndola mayor de lo que cabría esperar si todos los encuestados fueran independientes. Sin embargo, como no sabemos cuál era la incertidumbre de MusicWatch sobre su estimación en primer lugar, es difícil saber si la dependencia entre los encuestados podría desempeñar un papel.

6.7 Las poblaciones tienen muchas formas

Hay una variedad de estrategias que se pueden emplear para establecer un marco formal para hacer declaraciones inferenciales. A menudo, hay literalmente una población de unidades (por ejemplo, personas, pingüinos, etc.) sobre la que se quieren hacer declaraciones. En esos casos, está claro de dónde procede la incertidumbre (muestreo de la población) y qué es exactamente lo que se intenta estimar (alguna característica de la población). Sin embargo, en otras aplicaciones puede que no esté tan claro qué

exactamente es la población y qué es exactamente lo que estás tratando de estimar. En esos casos, tendrás que ser más explícito a la hora de definir la población porque puede haber más de una posibilidad.

Series temporales

Algunos procesos se miden en el tiempo (cada minuto, cada día, etc.). Por ejemplo, podemos estar interesados en analizar datos consistentes en el precio de cierre diario de las acciones de Apple para el año natural 2014. Si quisiéramos hacer una inferencia a partir de este conjunto de datos, ¿cuál sería la población? Hay algunas posibilidades.

1. Podríamos argumentar que el año 2014 fue muestreado aleatoriamente de la población de *todos los años posibles* de datos, por lo que las inferencias que hacemos se aplican a otros años del precio de las acciones.
2. Podríamos decir que las acciones de Apple representan una muestra de *todo* el *mercado de valores*, por lo que podemos hacer inferencias sobre *otras acciones* a partir de este conjunto de datos.

Independientemente de lo que elija, es importante dejar claro a qué población se refiere antes de intentar hacer una inferencia a partir de los datos.

Procesos naturales

Los fenómenos naturales, como los terremotos, los incendios, los huracanes, los fenómenos meteorológicos y otros acontecimientos que se producen en la naturaleza, suelen registrarse a lo largo del tiempo y del espacio. En el caso de las mediciones puramente temporales, podríamos definir la población de la misma manera que definimos la población anteriormente con el ejemplo de las series temporales. Sin embargo, podemos tener datos que

sólo se mide en el espacio. Por ejemplo, podemos tener un mapa de los epicentros de todos los terremotos que se han producido en una zona. Entonces, ¿cuál es la población? Un enfoque común es decir que existe un *proceso estocástico no observado* que deja caer aleatoriamente terremotos en la zona y que nuestros datos representan una muestra aleatoria de este proceso. En ese caso, estamos utilizando los datos para intentar aprender más sobre este proceso no observado.

Datos como población

Una técnica que siempre es posible, pero que no se utiliza habitualmente, es tratar el conjunto de datos como una población. En este caso, no hay inferencia porque no hay muestreo. Como el conjunto de datos *es* la población, no hay incertidumbre sobre ninguna característica de la población. Puede que esto no parezca una estrategia útil, pero hay circunstancias en las que puede utilizarse para responder a preguntas importantes. En particular, hay ocasiones en las que no nos importan las cosas que están fuera del conjunto de datos.

Por ejemplo, es habitual en las organizaciones analizar los datos salariales para asegurarse de que las mujeres no cobran menos que los hombres por un trabajo comparable o de que no hay grandes desequilibrios entre los empleados de diferentes grupos étnicos. En este caso, las diferencias salariales entre los distintos grupos pueden calcularse en el conjunto de datos y se puede ver si las diferencias son lo suficientemente grandes como para ser preocupantes. La cuestión es que los datos responden directamente a una pregunta de interés, que es "¿Existen grandes diferencias salariales que deban abordarse?". En este caso no hay necesidad de hacer una inferencia sobre los empleados de fuera de la organización (no hay ninguno, por definición) o a los empleados de otras organizaciones sobre las que no se tendría ningún control. El conjunto de datos es la población y las respuestas a cualquier pregunta relacionada con la población están en

ese conjunto de datos.

7. Modelado formal

Este capítulo suele ser la parte del libro de texto o del curso de estadística en la que la gente tiende a toparse con un muro. En particular, suele haber muchas matemáticas. Las matemáticas son buenas, pero las matemáticas gratuitas no son buenas. No estamos a favor de eso.

Es importante darse cuenta de que a menudo es útil representar un modelo utilizando una notación matemática porque es una notación compacta y puede ser fácil de interpretar una vez que te acostumbras a ella. Además, escribir un modelo estadístico utilizando una notación matemática, en contraposición al lenguaje natural, te obliga a ser preciso en tu descripción del modelo y en tu declaración de lo que estás tratando de lograr, como la estimación de un parámetro.

7.1 ¿Cuáles son los objetivos del modelado formal?

Un objetivo clave de la modelización formal es desarrollar una especificación precisa de su pregunta y de cómo pueden utilizarse sus datos para responder a esa pregunta. Los modelos formales permiten identificar claramente lo que se intenta inferir de los datos y la forma que adoptan las relaciones entre las características de la población. Puede ser difícil lograr este tipo de precisión utilizando sólo palabras.

Los parámetros desempeñan un papel importante en muchos modelos estadísticos formales (en lenguaje estadístico, se conocen como *modelos estadísticos paramétricos*). Son números que utilizamos para representar características o asociaciones que existen en la población.

Dado que representan características de la población, los parámetros se consideran generalmente desconocidos, y nuestro objetivo es estimarlos a partir de los datos que recogemos.

Por ejemplo, supongamos que queremos evaluar la relación entre el número de onzas de refresco que consume un individuo al día y el IMC de esa persona. La pendiente de una línea que podría trazar visualizando esta relación es el parámetro que quiere estimar para responder a su pregunta: "¿Cuánto se espera que aumente el IMC por cada onza adicional de refresco consumida?" Más concretamente, estás utilizando un *modelo de regresión lineal* para formular este problema.

Otro objetivo de la modelización formal es desarrollar un marco riguroso con el que poder cuestionar y probar los resultados primarios. En este punto del análisis de los datos, has formulado y refinado tu pregunta, has explorado los datos visualmente y quizá hayas realizado algún modelo exploratorio. Lo más importante es que probablemente tenga una idea bastante clara de cuál es la respuesta a su pregunta, pero tal vez tenga algunas dudas sobre si sus conclusiones se mantendrán bajo un intenso escrutinio. Suponiendo que siga interesado en avanzar con sus resultados, aquí es donde la modelización formal puede desempeñar un papel importante.

7.2 Marco general

Podemos aplicar el epiciclo básico de análisis a la parte de modelización formal del análisis de datos. Seguimos queriendo establecer expectativas, recopilar información y refinar nuestras expectativas basándonos en los datos. En este contexto, estas tres fases tienen el siguiente aspecto.

1. **Fijar las expectativas**. La fijación de las expectativas consiste en desarrollar un *modelo primario* que represente

su mejor sentido de lo que proporciona la respuesta a su pregunta. Este modelo se elige en función de la información de que dispongas en ese momento.

2. **Recogida de información**. Una vez establecido el modelo primario, querremos crear un conjunto de modelos secundarios que cuestionen el modelo primario de alguna manera. A continuación veremos ejemplos de lo que esto significa.
3. **Revisar las expectativas**. Si nuestros modelos secundarios consiguen cuestionar nuestro modelo primario y ponen en duda las conclusiones del modelo primario, puede que tengamos que ajustar o modificar el modelo primario para reflejar mejor lo que hemos aprendido de los modelos secundarios.

Modelo primario

A menudo es útil empezar con un *modelo primario*. Este modelo probablemente se derivará de cualquier análisis exploratorio que ya haya realizado y servirá como candidato principal para algo que resuma sucintamente sus resultados y se ajuste a sus expectativas. Es importante tener en cuenta que, en cualquier momento del análisis de datos, el modelo primario *no* es *necesariamente el modelo final*. Es simplemente el modelo con el que se compararán otros modelos secundarios. El proceso de comparar su modelo con otros modelos secundarios suele denominarse *análisis de sensibilidad*, ya que le interesa ver la sensibilidad de su modelo a los cambios, como la adición o eliminación de predictores o la eliminación de valores atípicos en los datos.

A lo largo del proceso iterativo de modelización formal, es posible que decida que un modelo diferente es más adecuado como modelo principal. Esto está bien, y forma parte del proceso de establecer expectativas, recoger información y refinar las expectativas en función de los datos.

Modelos secundarios

Una vez que se haya decidido por un modelo primario, se suele desarrollar una serie de modelos secundarios. El objetivo de estos modelos es poner a prueba la legitimidad y solidez de su modelo primario y generar potencialmente pruebas contra su modelo primario. Si los modelos secundarios consiguen generar pruebas que refutan las conclusiones del modelo primario, es posible que tenga que revisar el modelo primario y comprobar si sus conclusiones siguen siendo razonables.

7.3 Análisis de asociación

Los análisis asociativos son aquellos en los que se busca una asociación entre dos o más características en presencia de otros factores potencialmente confusos. Hay tres clases de variables en las que es importante pensar en un análisis asociativo.

1. **Resultado**. El resultado es la característica del conjunto de datos que se cree que cambia junto con el **predictor clave**. Incluso si no está planteando una pregunta causal o mecanística, por lo que no cree necesariamente que el resultado *responda* a los cambios en el predictor clave, sigue siendo necesario definir un resultado para la mayoría de los enfoques de modelización formal.
2. **Predictor clave**. A menudo, en los análisis asociativos hay un predictor clave de interés (puede haber varios). Queremos saber cómo cambia el resultado con este predictor clave. Sin embargo, nuestra comprensión de esa relación puede verse dificultada por la presencia de posibles factores de confusión.

3. **Posibles factores de confusión**. Se trata de una gran clase de predictores que están relacionados tanto con el predictor clave como con el resultado. Es importante conocer bien cuáles son y si están disponibles en el conjunto de datos. Si un factor de confusión clave no está disponible en el conjunto de datos, a veces habrá un sustituto que esté relacionado con ese factor de confusión clave y que pueda ser sustituido.

Una vez que haya identificado estas tres clases de variables en su conjunto de datos, puede empezar a pensar en el modelado formal en un entorno asociativo.

La forma básica de un modelo en un análisis asociativo será

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

donde

- y es el resultado
- x es el predictor clave
- z es un posible factor de confusión
- ε es un error aleatorio independiente
- α es el intercepto, es decir, el valor y cuando $x = 0$ y $z = 0$
- β es el cambio en y asociado a un aumento de 1 unidad x, ajustando para z
- γ es el cambio en y asociado a un aumento de 1 unidad en z, ajustado por x

Se trata de un modelo lineal, y nuestro principal interés es estimar el coeficiente β, que cuantifica la relación entre el predictor clave x y el resultado y.

Aunque tendremos que estimar α y γ como parte del proceso de estimación de β, no nos importa realmente el

En la literatura estadística, coefi- cientes como α y γ se denominan a veces *parámetros molestos* porque tenemos que utilizar los datos para estimarlos y completar la especificación del modelo, pero no nos importa su valor.

El modelo anterior podría considerarse como el modelo principal. Hay un predictor clave y un factor de confusión en el modelo en el que quizás se sabe que hay que tener en cuenta ese factor de confusión. Este modelo puede producir resultados sensatos y sigue lo que se conoce generalmente en el área.

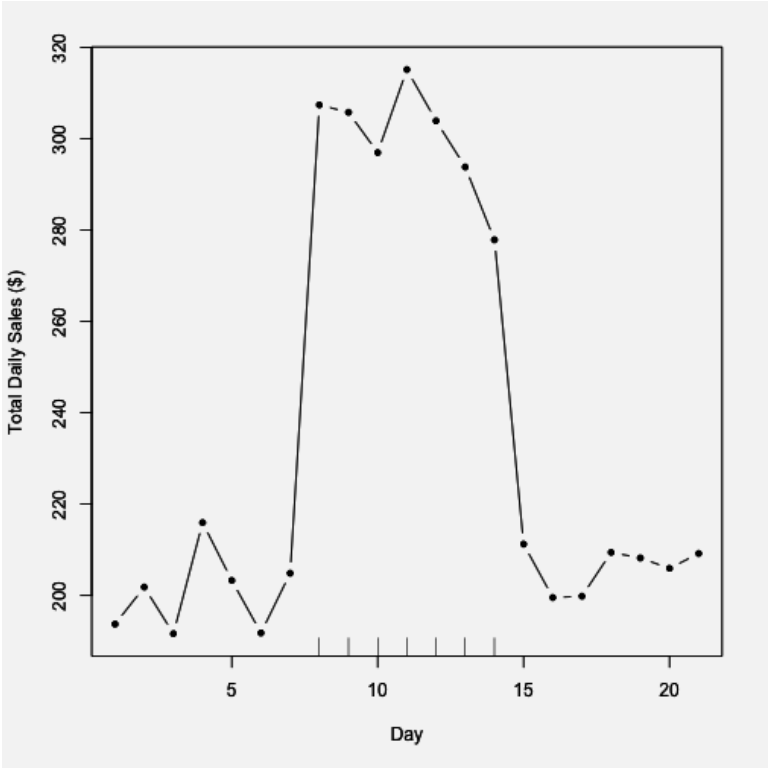
Ejemplo: Campaña de publicidad en línea

Supongamos que vendemos un nuevo producto en la web y nos interesa saber si la compra de anuncios en Facebook ayuda a aumentar las ventas de ese producto. Para empezar, podríamos iniciar una campaña publicitaria piloto de una semana en Facebook y medir el éxito de esa campaña. Si tiene éxito, podríamos seguir comprando anuncios para el producto.

Un enfoque sencillo podría ser hacer un seguimiento de las ventas diarias antes, durante y después de la campaña publicitaria (tenga en cuenta que hay formas más precisas de hacer esto con URL de seguimiento y Google Analytics, pero dejémoslo de lado por ahora). En pocas palabras, si la campaña tuviera una duración de una semana, podríamos observar la semana anterior, la semana durante y la semana posterior para ver si hubo algún cambio en las ventas diarias.

Expectativas

En un mundo ideal, los datos podrían ser algo así.



Campaña publicitaria hipotética

Las marcas en el eje de las abscisas indican el periodo en que la campaña estuvo activa. En este caso, es bastante obvio el efecto que tuvo la campaña publicitaria en las ventas. Sólo con los ojos, se puede decir que la campaña publicitaria añadió unos
100 dólares al día al total de ventas diarias. Su modelo principal podría ser algo así

$$y = \alpha + \beta x + \varepsilon$$

donde y es el total de ventas diarias y x es un indicador de si un día determinado cayó durante la campaña publicitaria o no. El hipo-

Los datos políticos para el gráfico anterior podrían tener el siguiente aspecto.

```
      campaña de ventas
día 1
                  193.7
35501
2                     201.836402
3                     191.643703
4                     215.952804
5                     203.295105
6                     191.795306
7                     204.874307
8                     307.383218
9                     305.757819
10                296.9461110
11                315.1178111
12                303.8984112
13                293.7876113
14                277.8530114
15                211.2493015
16                199.5507016
17                199.8381017
18                209.4384018
19                208.2122019
20                205.9390020
21                209.1898021
```

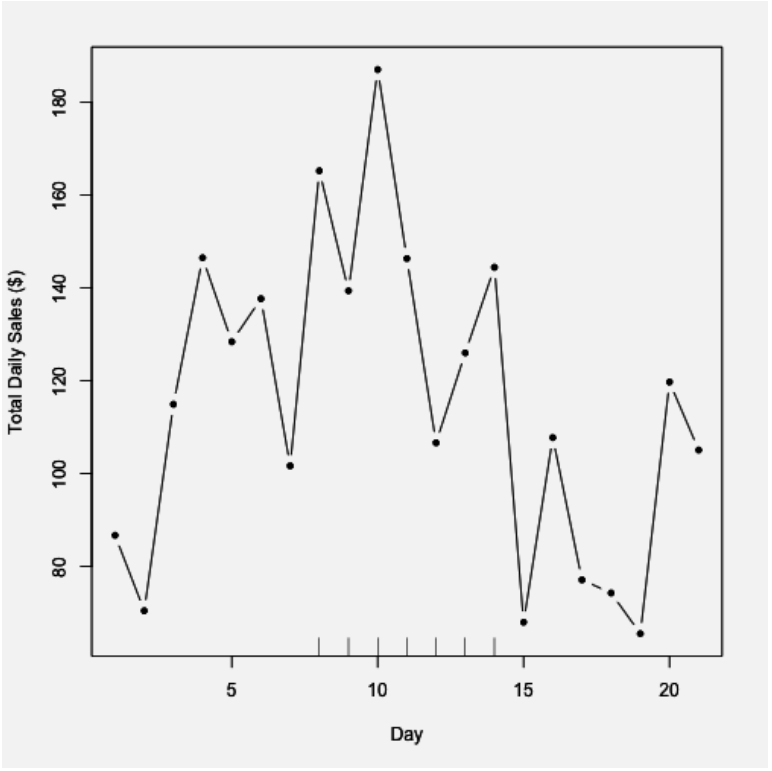
Teniendo en cuenta estos datos y el modelo primario anterior, estimamos que β es de 96,78 dólares, lo que no está muy lejos de nuestra estimación original de 100 dólares.

Establecimiento de expectativas. La discusión de este escenario ideal es importante no porque sea probable que ocurra, sino porque nos enseña lo que *esperaríamos* ver si el mundo funcionara de acuerdo con un marco más simple y cómo analizaríamos los datos bajo esas expectativas.

Datos más realistas

Por desgracia, rara vez vemos datos como los del gráfico anterior. En la realidad, los tamaños del efecto tienden a ser más pequeños, el ruido tiende a

ser más alto, y suele haber otros factores en juego. Normalmente, los datos serán algo así.



Datos de ventas diarias más realistas

Aunque parece que hay un aumento de las ventas durante el periodo de la campaña publicitaria (indicado por las marcas de nuevo), es un poco difícil argumentar que el aumento de las ventas *se* debe a la campaña. De hecho, en los días anteriores al inicio de la campaña, parece haber un ligero aumento de las ventas. ¿Es una casualidad o hay otras tendencias en el fondo? Es posible que haya una tendencia de fondo suave, de modo que las ventas diarias tienden a subir y bajar a lo largo del mes. Por lo tanto, incluso sin la campaña publicitaria en marcha,

es posible que hubiéramos visto un aumento de las ventas de todos modos. La cuestión ahora es si la campaña publicitaria aumentó las ventas diarias *además de* esta tendencia de fondo existente.

Tomemos nuestro modelo principal, que sólo incluye el resultado y el indicador de nuestra campaña publicitaria como predicción clave. Utilizando ese modelo, estimamos que β, el aumento de las ventas diarias debido a la campaña publicitaria, es de 44,75 dólares.

Sin embargo, supongamos que incorporamos una tendencia de fondo a nuestro modelo, por lo que en lugar de nuestro modelo principal, ajustamos el siguiente.

$$y = \alpha + \beta x + \gamma_1 t + \gamma_2 t^2 + \varepsilon$$

donde *t* indica ahora el número de día (es decir, 1, 2, ... , 21). Lo que hemos hecho es añadir una función cuadrática de *t* al modelo para permitir cierta curvatura en la tendencia (a diferencia de una función lineal que sólo permitiría un patrón estrictamente creciente o decreciente). Utilizando este modelo estimamos que β es de 39,86 dólares, que es algo menos que lo que el modelo primario estimaba para β.

Podemos ajustar un último modelo, que permite una tendencia de fondo aún más flexible: utilizamos un polinomio de 4º orden
para representar esa tendencia. Aunque nuestro modelo cuadrático nos parezca lo suficientemente complejo, el propósito de este último modelo es simplemente ampliar un poco los límites para ver cómo cambian las cosas en circunstancias más extremas. Este modelo nos da una estimación de β de 49,1 dólares, que de hecho es mayor que la estimación de nuestro modelo principal.

En este punto tenemos un modelo primario y dos modelos secundarios, que dan estimaciones algo diferentes de la asociación entre nuestra campaña publicitaria y las ventas totales diarias.

Modelo	Características	Estimación para β
Modelo 1 (primario)	Sin factores de confusión	$44.75
Modelo 2 (secundario)	Tendencia temporal cuadrática	$39.86
Modelo 3 (secundario)	Tiempo de la 4ª orden tendencia	$49.1

Evaluación

La determinación del camino a seguir puede depender de factores ajenos al conjunto de datos. Algunas consideraciones típicas son

1. **Tamaño del efecto**. Los tres modelos presentan un rango de es- tados de 39,86 a 49,1 dólares. ¿Es un rango grande? Es posible que para su organización un rango de esta magnitud no sea lo suficientemente grande como para hacer una diferencia y, por lo tanto, todos los modelos podrían considerarse equivalentes. O puede que considere que estas 3 estimaciones son significativamente diferentes entre sí, en cuyo caso podría dar más peso a un modelo que a otro. Otro factor podría ser el coste de la campaña publicitaria, en cuyo caso le interesaría el rendimiento de su inversión en los anuncios. Un aumento de 39,86 dólares al día podría valer la pena si el coste total de los anuncios fuera de 10 dólares al día, pero quizá no si el coste fuera de 20 dólares al día. En ese caso, es posible que necesite que el aumento de las ventas sea mayor para que la campaña merezca la pena. La cuestión aquí es que hay algunas pruebas de su modelo formal que indican que la campaña publicitaria sólo podría aumentar sus ventas diarias totales en 39,86, sin

embargo, otras pruebas dicen que podría ser mayor. La cuestión es si cree que merece la pena el riesgo de comprar más anuncios,

dado el abanico de posibilidades, o si cree que incluso en el extremo superior, probablemente no merezca la pena.

2. **Plausibilidad**. Aunque puede ajustar una serie de modelos con el fin de cuestionar su modelo principal, puede darse el caso de que algunos modelos sean más plausibles que otros, en términos de estar cerca de lo que sea la "verdad" sobre la población. En este caso, el modelo con una tendencia cuadrática parece plausible porque es capaz de captar un posible patrón de subida y bajada en los datos, si es que lo hay. El modelo con el polinomio de 4º orden es igualmente capaz de capturar este patrón, pero parece demasiado complejo para caracterizar un patrón simple como ese. El hecho de que un modelo pueda considerarse más o menos plausible dependerá de sus conocimientos sobre el tema y de su capacidad para relacionar los acontecimientos del mundo real con la fórmula matemática del modelo. Es posible que tengas que consultar a otros expertos en la materia para evaluar la plausibilidad de varios modelos.
3. **Parsimonia**. En el caso de que los diferentes modelos cuenten todos la misma historia (es decir, las estimaciones de β están lo suficientemente cerca como para ser consideradas "iguales"), a menudo es preferible elegir el modelo más sencillo. Hay dos razones para ello. En primer lugar, con un modelo más sencillo puede ser más fácil contar una historia sobre lo que ocurre en los datos a través de los distintos parámetros del modelo. Por ejemplo, es más fácil explicar una tendencia lineal que una tendencia exponencial. En segundo lugar, los modelos más sencillos, desde el punto de vista estadístico, son más "eficientes", es decir, aprovechan mejor los datos por cada parámetro que se estima. La complejidad de un modelo estadístico se refiere generalmente al número de parámetros del modelo; en este ejemplo, el modelo principal tiene 2 parámetros, mientras que el modelo más complejo tiene 6 parámetros. Si ningún

modelo

produce mejores resultados que otro, podríamos preferir un modelo que sólo contenga 2 parámetros porque es más sencillo de describir y es más parsimonioso. Si los modelos primario y secundario producen diferencias significativas, podríamos elegir un modelo parsimonioso en lugar de un modelo más complejo, pero no si el modelo más complejo cuenta una historia más convincente.

7.4 Análisis de predicción

En la sección anterior hemos descrito los análisis asociativos, en los que el objetivo es ver si un predictor clave *x* y un resultado *y* están asociados. Pero a veces el objetivo es utilizar toda la información disponible para predecir *y*. Además, no importa si las variables se consideran no relacionadas de forma causal con el resultado que se quiere predecir, porque el objetivo es la predicción, no el desarrollo de un entendimiento sobre las relaciones entre características.

Con los modelos de predicción, tenemos variables de resultado -características sobre las que nos gustaría hacer predicciones- pero normalmente no hacemos una distinción entre "predictores clave" y otros predictores. En la mayoría de los casos, cualquier predictor que pueda ser útil para predecir el resultado se tendrá en cuenta en un análisis y, *a priori,* se le dará el mismo peso en términos de su importancia para predecir el resultado. Los análisis de predicción suelen dejar que el algoritmo de predicción determine la importancia de cada predictor y la forma funcional del modelo.

Para muchos análisis de predicción no es posible escribir literalmente el modelo que se utiliza para predecir porque no puede representarse utilizando la notación matemática estándar. Muchas rutinas de predicción modernas están estructuradas como algoritmos o procedimientos que toman entradas y trans

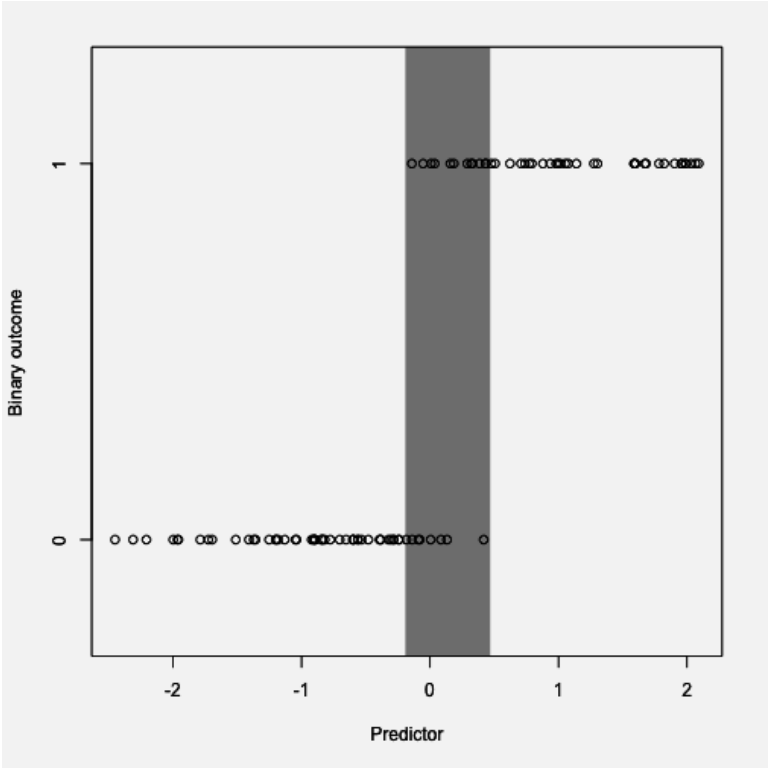
para convertirlos en resultados. El camino que siguen las entradas para transformarse en salidas puede ser altamente no lineal y los predictores pueden interactuar con otros predictores en el camino. Normalmente, no hay parámetros de interés que intentemos estimar; de hecho, muchos procedimientos algorítmicos no tienen ningún parámetro estimable.

La clave que hay que recordar en los análisis de predicción es que normalmente no nos importan los detalles específicos del modelo. En la mayoría de los casos, mientras el método "funcione", sea reproducible y produzca buenas predicciones con un error mínimo, habremos alcanzado nuestros objetivos.

En el caso de los análisis de predicción, el tipo exacto de análisis que se realiza depende de la naturaleza del resultado (como ocurre con todos los análisis). Los problemas de predicción suelen adoptar la forma de un **problema de clasificación en** el que el resultado es binario. En algunos casos, el resultado puede tener más de dos niveles, pero el caso binario es, con mucho, el más común. En esta sección, nos centraremos en el problema de clasificación binaria.

Expectativas

¿Cuál es el escenario ideal en un problema de predicción? Por lo general, lo que queremos es que un predictor, o un conjunto de predictores, produzca una *buena separación en* el resultado. Este es un ejemplo de un único predictor que produce una separación razonable en un resultado binario.



Escenario de clasificación ideal

El resultado toma valores de 0 y 1, mientras que el predictor es continuo y toma valores entre -2 y 2 aproximadamente. La zona gris indicada en el gráfico destaca el área en la que los valores del predictor pueden tomar valores de 0 o 1. A la derecha de la zona gris se observa que el valor del resultado es siempre 1 y a la izquierda de la zona gris el valor del resultado es siempre 0. En los problemas de predicción, es en esta zona gris donde tenemos la mayor incertidumbre sobre el resultado, dado el valor del predictor.

El objetivo de la mayoría de los problemas de predicción es identificar un conjunto de predictores que minimice el tamaño de esa zona gris en

el gráfico anterior. De forma contraria, es frecuente identificar predictores (especialmente categóricos) que *separan perfectamente* el resultado, de forma que la zona gris se reduce a cero. Sin embargo, estas situaciones suelen indicar un problema degenerado que no tiene mucho interés o incluso un error en los datos. Por ejemplo, una variable continua que ha sido dicotomizada estará perfectamente separada por su contraparte continua. Es un error común incluir la versión continua como predictor en el modelo y la versión dicotómica como resultado. En los datos del mundo real, se puede ver una separación casi perfecta cuando se miden rasgos o características que se sabe que están vinculados entre sí de forma mecánica o a través de algún proceso determinista. Por ejemplo, si el resultado fuera un indicador del potencial de una persona para contraer cáncer de ovario, el sexo de la persona podría ser un muy buen predictor, pero no es probable que sea uno de gran interés para nosotros.

Datos del mundo real

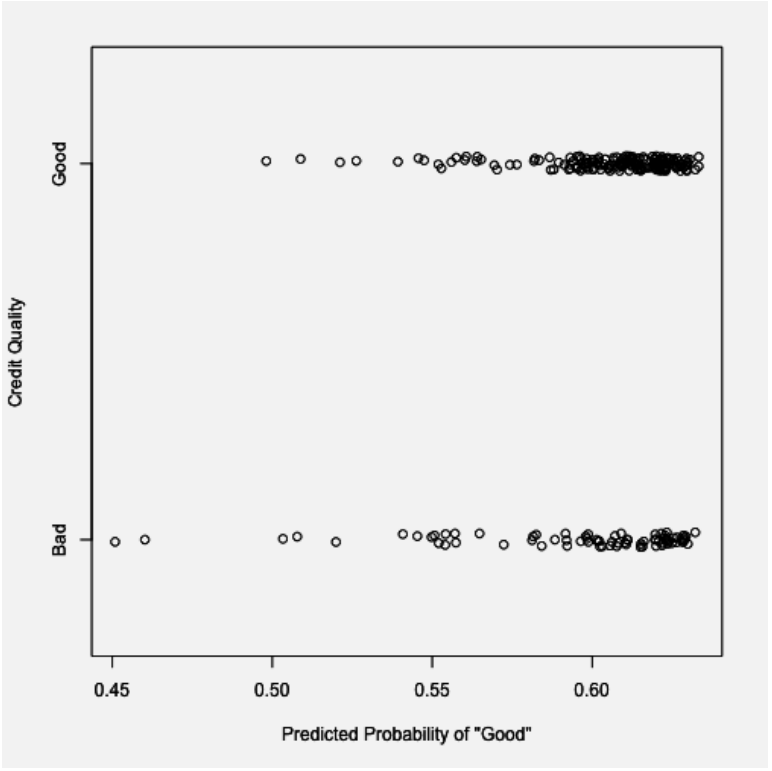
Para este ejemplo utilizaremos datos sobre la solvencia de los individuos. El conjunto de datos se ha extraído del Repositorio de Aprendizaje de Maquinaria de la UCI[1]. El conjunto de datos clasifica a los individuos en riesgos crediticios "buenos" o "malos" e incluye una variedad de predictores que pueden predecir la solvencia crediticia. Hay un total de 1.000 observaciones en el conjunto de datos y 62 características. A efectos de esta exposición, omitimos el código de este ejemplo, pero los archivos de código pueden obtenerse en el sitio web del Libro.

Lo primero que hacemos en un problema de predicción es dividir los datos en un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El conjunto de datos de entrenamiento sirve para desarrollar y ajustar el modelo y el conjunto de datos de prueba sirve para evaluar nuestro modelo ajustado y

[1] https://archive.ics.uci.edu/ml/datasets/Statlog+(Alemán+Crédito+Datos)

estimar su tasa de error. En este ejemplo utilizamos un 75% aleatorio de las observaciones para que sirvan de conjunto de datos de entrenamiento. El 25% restante servirá como conjunto de datos de prueba.

Tras ajustar el modelo al conjunto de datos de entrenamiento, podemos calcular las probabilidades previstas de tener un crédito "bueno" a partir del conjunto de datos de prueba. A continuación, representamos esas probabilidades predichas en el eje de abscisas junto con el estado crediticio real de cada individuo en el eje de ordenadas. (Las coordenadas del eje Y se han modificado aleatoriamente para mostrar más detalles).



Predicción frente a la verdad

Aquí podemos ver que no hay una buena separación

que vimos en el escenario ideal. En toda la gama de probabilidades predichas, hay personas con crédito "bueno" y "malo". Esto sugiere que el algoritmo de predicción que hemos empleado quizá tenga dificultades para encontrar una buena combinación de características que pueda separar a las personas con riesgo de crédito bueno y malo.

A continuación podemos calcular algunas estadísticas resumidas sobre el algoritmo de predicción.

```
Matriz de confusión y referencia

          estadística
Predicción Mala Buena
      Malas2      1
      Bien 73   174

               Precisión : 0.704
            IC DEL 95% : (0.6432, 0.7599)
   Sin información Tasa : 0.7
    Valor P [Acc > NIR] : 0.4762

                  Kappa : 0.0289
  Prueba de Mcnemar Valor <2e-16
                      P :

            Sensibilidad : 0.99429
          Especificidad : 0.02667
        Valor Pos Pred : 0.70445
        Valor Neg Pred : 0.66667
           Prevalencia : 0.70000
      Tasa de detección : 0.69600
        Prevalencia de la 0.98800
              detección :
  Precisión equilibrada : 0.51048

       Clase "positiva" : Buena
```

Podemos ver que la precisión es de aproximadamente un 70%, lo que no es muy bueno para la mayoría de los algoritmos de predicción. En particular, la especificidad del al- goritmo es muy pobre, lo que significa que si

"Riesgo de crédito malo", la probabilidad de que se le clasifique como tal es sólo de un 2,6%.

Evaluación

En el caso de los problemas de predicción, la decisión sobre el siguiente paso tras el ajuste inicial del modelo puede depender de algunos factores.

1. **Calidad de la predicción**. ¿Es la precisión del modelo lo suficientemente buena para sus objetivos? Esto depende del objetivo final
   y los riesgos asociados a las acciones posteriores. En el caso de las aplicaciones médicas, en las que el resultado puede ser la presencia de una enfermedad, es posible que queramos tener una alta sensibilidad, de modo que si realmente se tiene la enfermedad, el algoritmo la detecte. De este modo, podremos ponerle en tratamiento rápidamente. Sin embargo, si el tratamiento es muy doloroso, tal vez con muchos efectos secundarios, entonces podríamos preferir una alta especificidad, lo que garantizaría que no tratamos por error a alguien que *no* tiene la enfermedad. En el caso de las aplicaciones financieras, como el ejemplo de la solvencia crediticia utilizado aquí, puede haber costes asimétricos asociados a confundir un buen crédito con uno malo, frente a un mal crédito con uno bueno.
2. **Ajuste del modelo**. Un rasgo distintivo de los algoritmos de predicción son sus numerosos parámetros de ajuste. A veces estos parámetros pueden tener grandes efectos en la calidad de la predicción si se cambian, por lo que es importante estar informado del impacto de los parámetros de ajuste para cualquier al- goritmo que se utilice. No hay ningún algoritmo de predicción para el que un único conjunto de parámetros de ajuste funcione bien para todos los problemas. Lo más probable es que, para el ajuste inicial del modelo, utilice los

parámetros "por defecto", pero estos valores por defecto pueden no ser suficientes para sus propósitos. Si juega con los parámetros de ajuste, la calidad de la predicción puede cambiar en gran medida.

sus predicciones. Es muy importante que documente los valores de estos parámetros de ajuste para que el análisis pueda reproducirse en el futuro.

3. **Disponibilidad de otros datos**. Muchos algoritmos de predicción son bastante buenos a la hora de explorar la estructura de conjuntos de datos grandes y complejos y de identificar una estructura que pueda predecir mejor el resultado. Si descubre que su modelo no funciona bien, incluso después de ajustar algunos parámetros, es probable que necesite datos adicionales para mejorar su predicción.

7.5 Resumen

La modelización formal suele ser el aspecto más técnico del análisis de datos, y su propósito es establecer con precisión cuál es el objetivo del análisis y proporcionar un marco riguroso para cuestionar los hallazgos y poner a prueba las suposiciones. El enfoque que se adopte puede variar en función de si la pregunta se refiere fundamentalmente a la estimación de una asociación para desarrollar una buena predicción.

8. Inferencia vs. Predicción: Implicaciones para la estrategia de modelización

Entender si se está respondiendo a una pregunta inferencial o a una pregunta de predicción es un concepto importante porque el tipo de pregunta que está respondiendo puede influir en gran medida en la estrategia de modelización que siga. Si no entiende claramente qué tipo de pregunta está formulando, puede acabar utilizando el tipo de enfoque de modelización equivocado y, en última instancia, sacar conclusiones erróneas de sus datos. El propósito de este capítulo es mostrarle lo que puede ocurrir cuando confunde una pregunta con otra.

Los puntos clave que hay que recordar son

1. En el caso de las **preguntas inferenciales,** el objetivo suele ser estimar una asociación entre un predictor de interés y el resultado. Por lo general, sólo hay un puñado de predictores de interés (o incluso sólo uno), sin embargo, suele haber muchas variables de confusión potenciales que hay que tener en cuenta. El objetivo principal de la modelización es estimar una asociación asegurándose de que se ajustan adecuadamente los posibles factores de confusión. A menudo, se realizan análisis de sensibilidad para ver si las asociaciones de interés son robustas a diferentes conjuntos de factores de confusión.
2. En las **preguntas de predicción**, el objetivo es identificar el modelo que *mejor predice* el resultado. Por lo general, no damos ninguna importancia *a priori* a los predictores, siempre que sean buenos para

predecir el resultado.

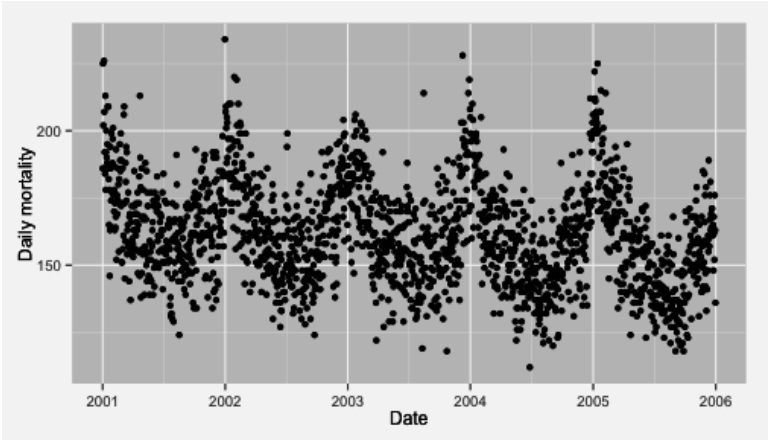
No existe la noción de "factor de confusión" o "predictores de interés" porque todos los predictores son potencialmente útiles para predecir el resultado. Además, a menudo no nos preocupamos de "cómo funciona el modelo" ni de contar una historia detallada sobre los predictores. El objetivo principal es desarrollar un modelo con una buena capacidad de predicción y estimar una tasa de error razonable a partir de los datos.

8.1 Contaminación del aire y mortalidad en la ciudad de Nueva York

El siguiente ejemplo muestra cómo diferentes tipos de preguntas y los correspondientes enfoques de modelización pueden llevar a diferentes conclusiones. El ejemplo utiliza datos de contaminación atmosférica y mortalidad de la ciudad de Nueva York. Los datos se utilizaron originalmente como parte del Estudio Nacional de Morbilidad, Mortalidad y Contaminación del Aire[1] (NMMAPS).

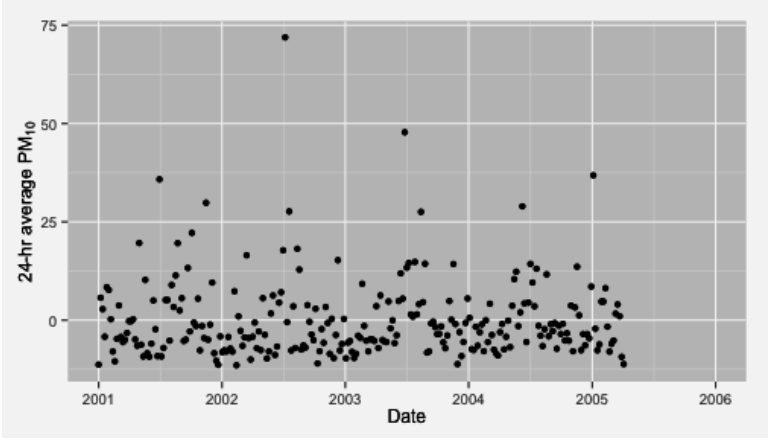
A continuación se muestra un gráfico de la mortalidad diaria por todas las causas para los años 2001-2005.

[1] http://www.ihapss.jhsph.edu



Mortalidad diaria en la ciudad de Nueva York, 2001-2005

Y aquí hay un gráfico de los niveles medios de 24 horas de partículas con diámetro aerodinámico inferior o igual a 10 micras (PM10).



PM10 diarias en la ciudad de Nueva York, 2001-2005

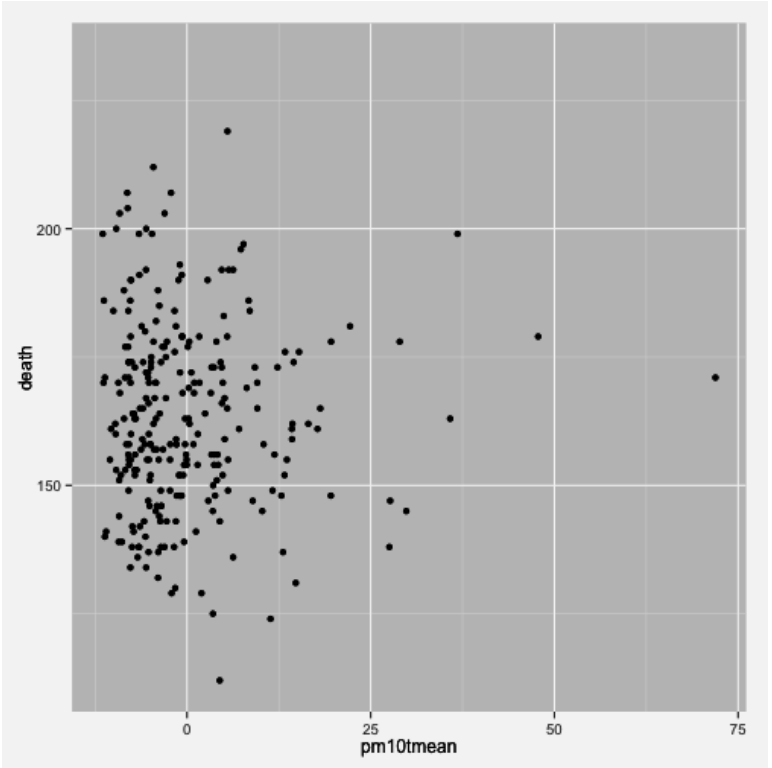
Obsérvese que hay muchos menos puntos en el gráfico anterior que en el de los datos de mortalidad. Esto se debe a que las PM10 no se miden todos los días. Obsérvese también que

hay valores negativos en el gráfico de PM10-esto se debe a que los datos de PM10 fueron sustraídos de la media. En general, los valores negativos de PM10 no son posibles.

8.2 Inferir una asociación

El primer enfoque que adoptaremos será preguntar "¿Existe una asociación entre los niveles medios diarios de PM10 en 24 horas y la mortalidad diaria?" Esta es una pregunta inferencial y estamos tratando de estimar una asociación. Además, para esta pregunta, sabemos que hay una serie de posibles factores de confusión con los que tendremos que lidiar.

Veamos la asociación bivariada entre las PM10 y la mortalidad. Aquí tenemos un gráfico de dispersión de las dos variables.



PM10 y mortalidad en la ciudad de Nueva York

No parece que haya mucho que hacer ahí, y un simple modelo de regresión lineal del logaritmo de la mortalidad diaria y las PM10 parece confirmarlo.

```
                              EstimaciónErrorortvalorPr(>|t|)
(Intercepción) 5,08884308354 0,0069353779 733,75138151
0,0000000
pm10media 0,00004033446 0,0006913941 0,05833786 0,9535247
```

En la tabla de coeficientes anterior, el coeficiente de `pm10tmean` es bastante pequeño y su error estándar es relativamente grande. Efectivamente, esta estimación de la asociación es cero.

Sin embargo, sabemos bastante sobre las PM10 y la mortalidad diaria, y una de las cosas que sabemos es que *la estación* desempeña unpapel importante en ambas variables. En particular, sabemos que la mortalidad tiende a ser mayor en invierno y menor en verano. Las PM10 tienden a mostrar el patrón inverso, siendo más altas en verano y más bajas en invierno. Dado que la estación está relacionada *tanto con las PM10* como con la mortalidad, es un buen candidato para ser un factor de confusión y tendría sentido ajustarlo en el modelo.

Aquí están los resultados de un segundo modelo, que incluye tanto las PM10 como la estación. La estación se incluye como una variable indicadora con 4 niveles.

```
                          EstimateStd.  Errort
                          valuePr(>|t|) (Intercept)
5.166484285 0.0112629532 458.714886 0.000000e+00
temporadaQ   -0.109271301 0.0166902948  -6.546996 3.209291e-10
2
temporadaQ   -0.155503242 0.0169729148  -9.161847 1.736346e-17
3
temporadaQ   -0.060317619 0.0167189714  -3.607735 3.716291e-04
4
pm10tmean     0.001499111 0.0006156902   2.434847 1.558453e-02
```

Observe ahora que el coeficiente `pm10tmean` es bastante mayor que antes y su `valor t` es grande, lo que sugiere una fuerte asociación. ¿Cómo es posible?

Resulta que tenemos un ejemplo clásico de la Paradoja de Simpson[2] de Simpson. La relación global entre P10 y la mortalidad es nula, pero cuando tenemos en cuenta la variación estacional tanto de la mortalidad como de las PM10, la asociación es positiva. El sorprendente resultado proviene de las formas opuestas en que la estación está relacionada con la mortalidad y las PM10.

Hasta ahora hemos tenido en cuenta la estación, pero hay otros factores de confusión potenciales. En particular, las variables meteorológicas, como la temperatura y la temperatura del punto de rocío, también están relacionadas con la formación de PM10 y la mortalidad.

https://en.wikipedia.org/wiki/Simpson%27s_paradox

En el siguiente modelo incluimos la temperatura (`tmpd`) y la temperatura del punto de rocío (`dptp`). También incluimos la variable de `la fecha` en caso de que haya alguna tendencia a largo plazo que deba tenerse en cuenta.

```
                              EstimateStd.  Errort
                              valuePr(>|t|) (Intercept)
5.62066568788 0.16471183741 34.1242365 1.851690e-96
fecha-0     .00002984198 0.00001315212 -2.2689856 2.411521e-02
temporadaQ2-0        ,05805970053 0,02299356287 -2,5250415 1,218288e-02
temporadaQ3-0        ,07655519887 0,02904104658 -2,6361033 8,906912e-03
temporadaQ4-0        ,03154694305 0,01832712585 -1,7213252 8,641910e-02
tmpd-0      .00295931276 0.00128835065 -2.2969777 2.244054e-02
dptp0        .00068342228 0.00103489541 0.6603781 5.096144e-01
pm10media0   .00237049992 0.00065856022 3.5995189 3.837886e-04
```

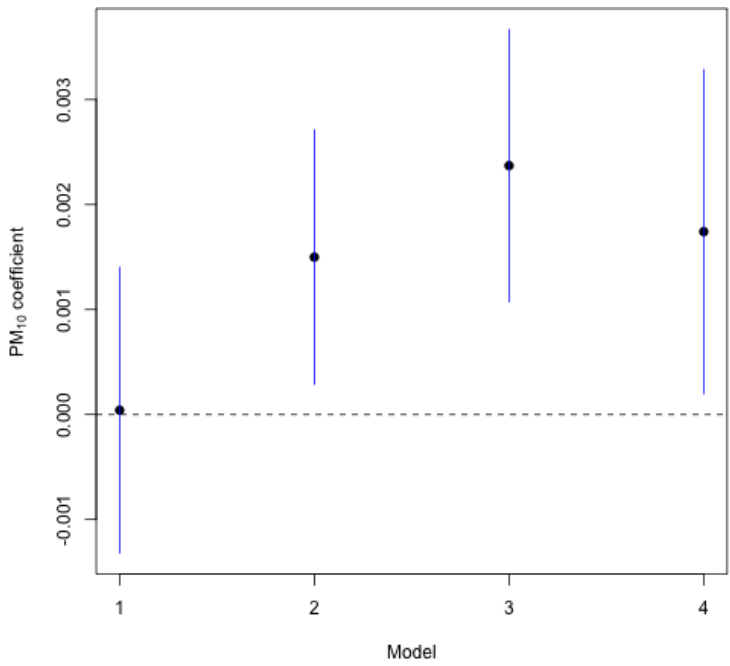
Obsérvese que el coeficiente `medio de PM10` es aún mayor que en el modelo anterior. Parece que sigue habiendo una asociación entre las PM10 y la mortalidad. El tamaño del efecto es pequeño, pero lo discutiremos más adelante.

Por último, otra clase de posibles factores de confusión son otros contaminantes. Antes de culpar a las PM10 como contaminante perjudicial, es importante que examinemos si puede haber otro contaminante que explique lo que estamos observando. El NO2 es un buen candidato porque comparte algunas de las mismas fuentes que las PM10 y se sabe que está relacionado con la mortalidad. Veamos qué ocurre cuando lo incluimos en el modelo.

	Estimación	Error estándar	valor t	Pr(>|t|)
(Intercepción)	5.61378604085	0.16440280471	34.1465345	2.548704e-96
fecha	-0.00002973484	0.00001312231	-2.2659756	2.430503e-02
temporadaQ2	-0.05143935218	0.02338034983	-2.2001105	2.871069e-02
temporadaQ3	-0.06569205605	0.02990520457	-2.1966764	2.895825e-02
temporadaQ4	-0.02750381423	0.01849165119	-1.4873639	1.381739e-01
tmpd	-0.00296833498	0.00128542535	-2.3092239	2.174371e-02
dptp	0.00070306996	0.00103262057	0.6808599	4.965877e-01
no2tmean	0.00126556418	0.00086229169	1.4676753	1.434444e-01
pm10tmean	0.00174189857	0.00078432327	2.2208937	2.725117e-02

Obsérvese en la tabla de coeficientes que el coeficiente `no2tmean` es de magnitud similar al coeficiente `pm10tmean`, aunque su `valor t` no es tan grande. El coeficiente `pm10tmean` parece ser estadísticamente significativo, pero su magnitud es ahora algo menor.

A continuación se muestra un gráfico del coeficiente de PM10 de los cuatro modelos que hemos probado.



Asociación entre las PM10 y la mortalidad según diferentes modelos

Con la excepción del modelo 1, que no tuvo en cuenta ningún factor de confusión potencial, parece haber una asociación positiva entre las PM10 y la mortalidad en todos los modelos 2-
4. Lo que esto signifique y lo que debamos hacer al respecto depende de cuál sea nuestro objetivo final y no lo discutiremos en detalle aquí. Es notable que el tamaño del efecto es generalmente pequeño, especialmente comparado con algunos de los otros predictores en el modelo. Sin embargo, también vale la pena señalar que, presumiblemente, todo el mundo en la ciudad de Nueva York respira, por lo que un pequeño efecto podría tener un gran impacto.

8.3 Predicción del resultado

Otra estrategia que podríamos haber adoptado es preguntar "¿Qué es lo que mejor predice la mortalidad en la ciudad de Nueva York?". Esta es claramente una pregunta de predicción y podemos utilizar los datos que tenemos a mano para construir un modelo. En este caso, utilizaremos la estrategia de modelización de bosques aleatorios[3] , que es un enfoque de aprendizaje automático que funciona bien cuando hay un gran número de predictores. Un tipo de resultado que podemos obtener del procedimiento de bosques aleatorios es una medida de la *importancia de las variables*. A grandes rasgos, esta medida indica la importancia de una variable determinada para mejorar la capacidad de predicción del modelo.

A continuación se muestra un gráfico de importancia de las variables, que se obtiene tras ajustar un modelo de bosque aleatorio. Los valores más grandes en el eje x indican una mayor importancia.

[3]https://en.wikipedia.org/wiki/Random_forest

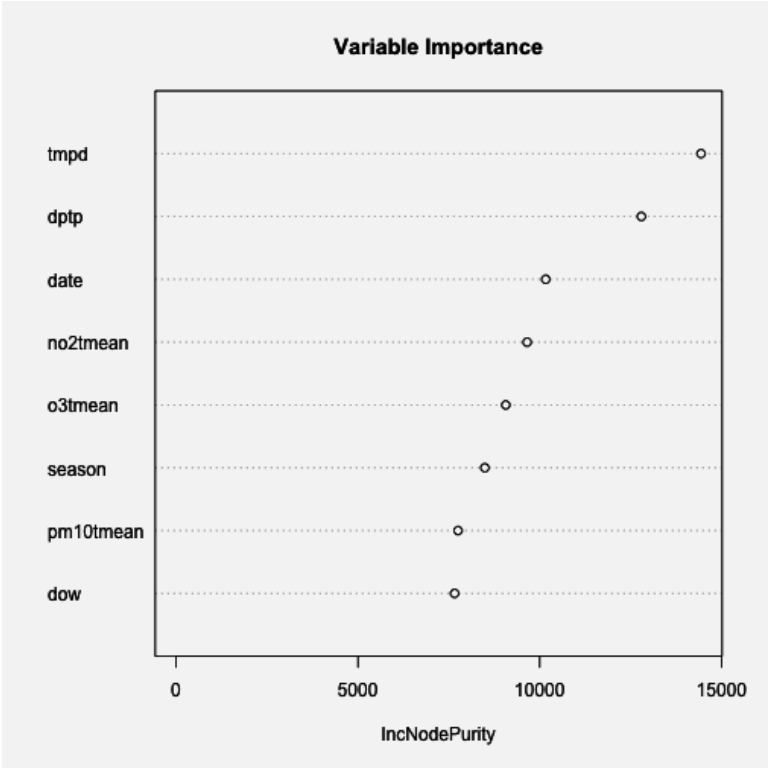


Gráfico de importancia de las variables de Random Forest para predecir la mortalidad

Obsérvese que la variable `pm10tmean` ocupa el último lugar de la lista en términos de importancia. Esto se debe a que no contribuye mucho a predecir el resultado, la mortalidad. Recordemos que en la sección anterior el tamaño del efecto parecía ser pequeño, lo que significa que no explicaba realmente mucha variabilidad en la mortalidad. Predictores como la temperatura y el punto de rocío son más útiles como predictores de la mortalidad diaria. Incluso el NO2 es un mejor predictor que las PM10.

Sin embargo, el hecho de que las PM10 no sean un fuerte predictor de la mortalidad no significa que no tengan una asociación relevante con la mortalidad. Teniendo en cuenta las compensaciones que han

que hacer al desarrollar un modelo de predicción, las PM10 no ocupan un lugar destacado en la lista de predictores que incluiríamos; sencillamente, no podemos incluir todos los predictores.

8.4 Resumen

En cualquier análisis de datos, hay que preguntarse: "¿Estoy haciendo una pregunta inferencial o una pregunta de predicción?". Esto debe aclararse *antes de* analizar los datos, ya que la respuesta a la pregunta puede guiar toda la estrategia de modelización. En este ejemplo, si nos hubiéramos decidido por un enfoque de predicción, podríamos haber pensado erróneamente que las PM10 no eran relevantes para la mortalidad. Sin embargo, el enfoque inferencial sugería una asociación estadísticamente significativa con la mortalidad. Plantear bien la cuestión y aplicar la estrategia de modelización adecuada puede influir mucho en el tipo de conclusiones que se extraigan de los datos.

9. Interpretación de los resultados

Aunque hemos dedicado un capítulo entero a la interpretación de los resultados de un análisis de datos, en realidad la interpretación ocurre continuamente a lo largo de un análisis. Es posible que los analistas de datos experimentados ni siquiera sean conscientes de la frecuencia con la que interpretan sus hallazgos porque se ha convertido en algo natural para ellos.

A estas alturas, el proceso epicicloidal de tres pasos: establecer expectativas, recopilar información (datos) y luego ajustar las expectativas a los datos, debería resultarle muy familiar, por lo que reconocerá que el tercer paso, ajustar las expectativas a los datos, es en sí mismo una interpretación. En cierto modo, hemos abordado el tema de la interpretación de los resultados a lo largo del libro. Sin embargo, merece su propio capítulo porque la interpretación es mucho más que la adecuación de las expectativas a los resultados y porque es, en sí misma, un paso importante del análisis de datos. Dado que la interpretación se realiza de forma más libre después de completar los análisis primarios y de apoyo, incluida la modelización formal, pero antes de comunicar los resultados, hemos colocado este capítulo entre estos capítulos respectivos.

9.1 Principios de interpretación

Existen varios principios de interpretación de resultados que ilustraremos en este capítulo. Estos principios son:

1. Revise su pregunta original

2. Comience con el modelo estadístico primario para orientarse y concéntrese en la naturaleza del resultado más que en una evaluación binaria del mismo (por ejemplo, estadísticamente significativo o no). La naturaleza del resultado incluye tres características: su direccionalidad, magnitud e incertidumbre. La incertidumbre es una evaluación de la probabilidad de que el resultado se haya obtenido por casualidad.
3. Desarrolle una interpretación global basada en (a) la totalidad de su análisis y (b) el contexto de lo que ya se conoce sobre el tema.
4. Considere las implicaciones, que le guiarán a la hora de determinar qué acción(es), en su caso, debería(n) tomarse como resultado de la respuesta a su pregunta.

Es importante señalar que el epiciclo del análisis también se aplica a la interpretación. En cada uno de los pasos de la interpretación, debe tener expectativas antes de realizar el paso, y luego ver si el resultado del paso coincide con sus expectativas. Sus expectativas se basan en lo que aprendió en el proceso de su análisis exploratorio de datos y de su modelización formal, y cuando su interpretación no coincide con sus expectativas, entonces tendrá que determinar si no coinciden porque sus expectativas son incorrectas o su interpretación es incorrecta. Aunque esté en uno de los últimos pasos del análisis de datos cuando interprete formalmente sus resultados, puede que tenga que volver al análisis exploratorio de datos o a la modelización para hacer coincidir las expectativas con los datos.

9.2 Estudio de caso: Refrescos no dietéticos Consumo e índice de masa corporal

Probablemente lo más fácil sea ver los principios de interpretación en acción para aprender a aplicarlos a su propia

El análisis de datos, por lo que utilizaremos un estudio de caso para ilustrar cada uno de los principios.

Volver a plantear la cuestión

El primer principio es recordar la pregunta original. Esto puede parecer una afirmación frívola, pero no es raro que la gente pierda el rumbo a medida que avanza en el proceso de análisis exploratorio y modelado formal. Esto suele ocurrir cuando un analista de datos se desvía demasiado del camino persiguiendo un hallazgo incidental que aparece en el proceso de análisis exploratorio de datos o de modelización formal. Entonces, el modelo o modelos finales proporcionan una respuesta a otra pregunta que surgió durante los análisis y no a la pregunta original.

Recordar su pregunta también sirve para proporcionar un marco para su interpretación. Por ejemplo, su pregunta original puede haber sido "Por cada lata de refresco de 12 onzas que se bebe al día, ¿cuánto aumenta el IMC medio entre los adultos de Estados Unidos?". La redacción de la pregunta le indica que su intención original era determinar cuánto mayor es el IMC entre los adultos de EE.UU. que beben, por ejemplo, dos latas de refrescos de 12 onzas al día de media, que entre los adultos que beben sólo un refresco de 12 onzas al día de media. La interpretación de sus análisis debería arrojar una afirmación como la siguiente Por cada lata de refresco de 12 onzas adicional que beben los adultos en EE.UU., el IMC aumenta, de media, en X kg/m^2 . Pero no debería producir una afirmación como: "Por cada *onza* adicional de refresco que beben los adultos en EE.UU., el IMC aumenta, de media, en X kg/m^2 ."

Otra forma en la que la revisión de la pregunta proporciona un marco para la interpretación de los resultados es que recordar el tipo de pregunta que se hizo proporciona

un marco explícito para la interpretación (véase Formular y afinar la pregunta para una revisión de los tipos de preguntas). Por ejemplo, si su pregunta fuera "Entre los adultos de EE.UU., ¿tienen los que beben 1 porción más de 12 onzas de refresco no dietético al día un IMC más alto, en promedio?", esto le indica que su pregunta es una pregunta *inferencial* y que su objetivo es comprender el efecto promedio de beber una porción adicional de 12 onzas de refresco no dietético al día sobre el IMC entre la población adulta de EE.UU.. Para responder a esta pregunta, es posible que haya realizado un análisis utilizando datos transversales recogidos en una muestra representativa de la población adulta de EE.UU., y en este caso su interpretación del resultado se enmarca en términos de cuál es la asociación entre una porción adicional de 12 onzas de refresco al día y el IMC, en promedio en la población adulta de EE.UU.

Dado que su pregunta no era causal y, por tanto, su análisis no era causal, el resultado no puede enmarcarse en términos de lo que ocurriría si una población empezara a consumir una lata más de refresco al día. Una pregunta causal podría ser: "¿Qué efecto tiene sobre el IMC beber una porción adicional de 12 onzas de refresco no dietético al día?", y para responder a esta pregunta, se podrían analizar los datos de un ensayo clínico que asignara aleatoriamente a un grupo a beber una lata adicional de refresco y al otro grupo a beber una lata adicional de una bebida placebo. Los resultados de este tipo de pregunta y análisis podrían interpretarse como cuál sería el efecto causal de beber una lata adicional de refresco de 12 onzas al día sobre el IMC. Dado que el análisis compara el efecto medio sobre el IMC entre los dos grupos (refresco y placebo), el resultado se interpretaría como el efecto causal medio en la población.

Un tercer objetivo de la revisión de la pregunta original es que es importante hacer una pausa y considerar si su enfoque para responder a la pregunta podría haber producido **una** re

resultado. Aunque ya hemos tratado el tema del sesgo en el capítulo sobre el planteamiento y el perfeccionamiento de la pregunta, a veces se adquiere nueva información durante el proceso de análisis exploratorio de los datos y/o la elaboración de modelos que afecta directamente a la evaluación de si el resultado puede estar sesgado. Recuerde que el sesgo es un problema sistemático con la recopilación o el análisis de los datos que da lugar a una respuesta incorrecta a su pregunta.

Utilizaremos el ejemplo de los refrescos y el IMC para ilustrar un ejemplo más sencillo de sesgo. Supongamos que su pregunta general sobre la relación entre los refrescos y el IMC hubiera incluido una pregunta inicial que fuera ¿Cuál es la media de consumo diario de refrescos no dietéticos entre los adultos de Estados Unidos? Supongamos que su análisis indica que en la muestra que está analizando, que es una muestra de todos los adultos de EE.UU., el número medio de porciones de 12 onzas de refrescos no dietéticos que se beben al día es de 0,5, por lo que infiere que el número medio de porciones de 12 onzas de refrescos que beben al día los adultos de EE.UU. también es de 0,5. Como siempre hay que cuestionar los resultados, es importante considerar si el análisis tiene un sesgo inherente.

¿Cómo se hace esto? Se empieza imaginando que el resultado es incorrecto, y luego se piensa en las formas en que la recopilación o el análisis de datos podría haber tenido un problema sistemático que dio lugar a una estimación incorrecta del número medio de latas de refresco no dietético de 12 onzas que beben al día los adultos en los EE.UU. Aunque este ejercicio de imaginar que su resultado es incorrecto se discute como un enfoque para evaluar el potencial de sesgo, esta es una excelente manera de **cuestionar sus resultados en cada paso del análisis**, ya sea que esté evaluando el riesgo de sesgo, o la confusión, o un problema técnico con su análisis.

El experimento mental es algo así: imagina que el número medio *real* de raciones de 12 onzas de productos no

Ahora imagine cómo el resultado de su análisis de la muestra, que fue de 0,5, podría estar tan alejado del resultado real: por alguna razón, la muestra de la población que comprende su conjunto de datos no es una muestra aleatoria de la población y, en cambio, tiene un número desproporcionado de personas que no beben ningún refresco no dietético, lo que hace bajar la media estimada de 12 onzas de porciones de refrescos no dietéticos consumidos al día. También podría imaginar que si el resultado de su muestra hubiera sido 4, que es mucho más alto que la cantidad real que beben al día los adultos en EE.UU., su muestra tiene un número desproporcionado de personas que tienen un alto consumo de refrescos no dietéticos, de modo que la estimación generada a partir de sus análisis es mayor que el valor real. Entonces, ¿cómo puede calibrar si su muestra no es aleatoria?

Para averiguar si su muestra es una muestra no aleatoria de la población objetivo, piense en lo que podría haber ocurrido para atraer a más personas que no consumen refrescos no dietéticos (o a más personas que consumen muchos) para ser incluidas en la muestra. Tal vez el estudio se anunció para la participación en una revista de fitness, y los lectores de revistas de fitness son menos propensos a beber refrescos no dietéticos. O tal vez los datos se recogieron mediante una encuesta por Internet y los encuestados por Internet son menos propensos a beber refrescos no dietéticos. O tal vez la encuesta recogió información sobre el consumo de refrescos no dietéticos proporcionando una lista de refrescos no dietéticos y pidiendo a los encuestados que indicaran cuáles habían consumido, pero la encuesta omitió Mountain Dew y Cherry Coke, de modo que las personas que beben principalmente estos refrescos no dietéticos fueron clasificadas como no consumidores de refrescos no dietéticos (o consumen menos de los que realmente consumen). Y así sucesivamente.

Aunque ilustramos el escenario más simple para el sesgo, que

Si bien es cierto que la estimación de la prevalencia o de la media puede ser sesgada, también lo es la estimación de la relación entre dos variables. Por ejemplo, los métodos de encuesta podrían sobremuestrear involuntariamente a las personas que no consumen refrescos no dietéticos y que tienen un IMC alto (como las personas con diabetes de tipo 2), de modo que el resultado indicaría (incorrectamente) que consumir refrescos no dietéticos no está asociado a tener un IMC más alto. La cuestión es que detenerse a realizar un experimento mental deliberado sobre las fuentes de sesgo es de vital importancia, ya que es realmente la única forma de evaluar el potencial de un resultado sesgado. Este experimento mental también debería llevarse a cabo cuando se está planteando y refinando la pregunta y también cuando se están realizando análisis exploratorios y modelos.

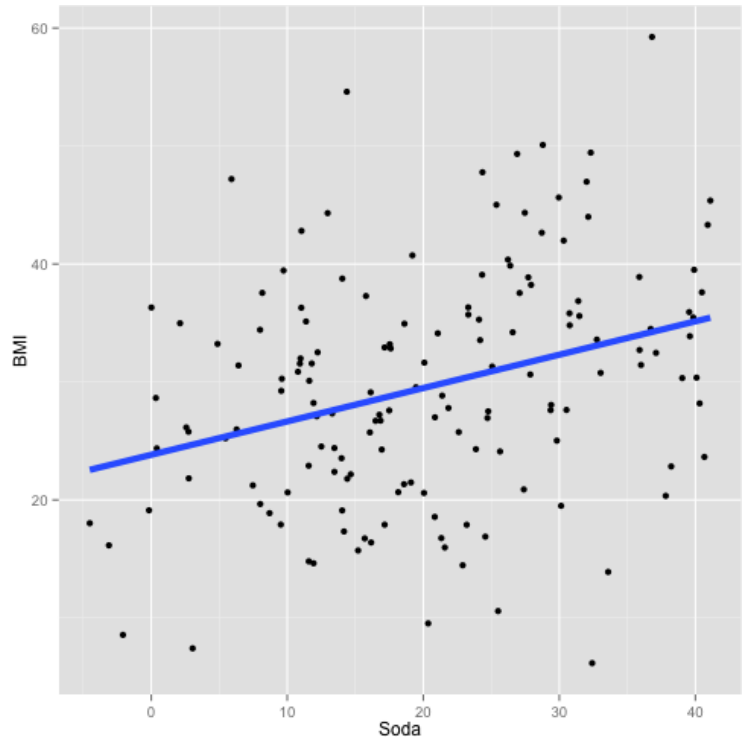
Comience con el modelo primario y evalúe la direccionalidad, magnitud e incertidumbre del resultado

El segundo principio es empezar con un único modelo y centrarse en todo el continuo del resultado, incluyendo su direccionalidad y magnitud, y el grado de certeza (o incertidumbre) que existe sobre si el resultado de la muestra que ha analizado refleja el verdadero resultado de la población global. Una gran cantidad de información necesaria para la interpretación de los resultados se perderá si se centra en una sola característica del resultado (como el valor p), de modo que ignora o pasa por alto otra información importante proporcionada por el modelo. Aunque su interpretación no está completa hasta que considere los resultados en su totalidad, a menudo es más útil centrarse primero en la interpretación de los resultados del modelo que cree que responde mejor a su pregunta y refleja (o "se ajusta") a sus datos, que es su modelo primario (Ver Modelado Formal). No dedique mucho tiempo a preocuparse por el modelo único con el que debe empezar, porque en

Al final, considerará todos sus resultados y este ejercicio de interpretación inicial le servirá para orientarse y proporcionar un marco para su interpretación final.

Direccionalidad

Basándonos en el ejemplo de la soda y el IMC, veamos el siguiente conjunto de datos de muestra con un modelo ajustado superpuesto.



Datos de muestra para el ejemplo de IMC-soda

Nos centraremos en lo que el modelo nos dice sobre la **direccionalidad** de la relación entre el consumo de refrescos y el IMC, la **magnitud** de la relación y la **incertidumbre.**

La relación entre el consumo de refrescos no dietéticos y el IMC es real y no un reflejo de la variación aleatoria que se espera al tomar una muestra de una población mayor.

El modelo indica que la direccionalidad de la relación es positiva, lo que significa que a medida que aumenta el consumo de refrescos no dietéticos, aumenta el IMC. Los otros resultados potenciales podrían haber sido una direccionalidad negativa, o ninguna direccionalidad (un valor de aproximadamente 0). ¿Coincide la direccionalidad positiva del resultado con sus expectativas desarrolladas a partir del análisis exploratorio de datos? Si es así, está en buena forma y puede pasar a la siguiente actividad de interpretación. Si no es así, hay un par de posibles explicaciones. En primer lugar, es posible que sus expectativas no sean correctas porque, o bien el análisis exploratorio se realizó de forma incorrecta, o bien su interpretación de los análisis exploratorios no fue correcta. En segundo lugar, el análisis exploratorio y su interpretación pueden ser correctos, pero la modelización formal puede haberse realizado de forma incorrecta. Tenga en cuenta que con este proceso está aplicando una vez más el epiciclo del análisis de datos.

Magnitud

Una vez que haya identificado y abordado cualquier discrepancia entre sus expectativas y la interpretación de la direccionalidad de la relación, el siguiente paso es considerar la **magnitud** de la relación. Como el modelo es una regresión lineal, puede ver que la pendiente de la relación, reflejada por el coeficiente beta, es de 0,28. La interpretación de la pendiente requiere conocer las unidades de la variable "refresco". Si las unidades son latas de refresco de 12 onzas al día, la interpretación de esta pendiente es que el IMC aumenta en 0,28 kg/m^2 por cada lata adicional de 12 onzas de refresco no dietético que se consuma al día. Sin embargo, las unidades están en onzas de refresco,

por lo que la interpretación de su modelo es que el IMC aumenta en 0,28 kg/m² por cada onza adicional de refresco no dietético que se consuma al día.

Aunque está seguro de que entiende las unidades de su variable de refresco correctamente y tiene la interpretación correcta del modelo, todavía no tiene la respuesta a su pregunta, que se formuló en términos de la asociación de cada lata adicional de 12 onzas de refresco y el IMC, no de cada onza adicional de refresco no dietético. Así que tendrá que convertir la pendiente de 0,28 para que se refiera a un aumento de 12 onzas, en lugar de 1 onza, en el consumo de refrescos. Dado que el modelo es un modelo lineal, basta con multiplicar la pendiente, o el coeficiente beta, por 12 para obtener 3,36, lo que indica que cada lata adicional de 12 onzas de refresco consumida al día está asociada a un IMC 3,36 kg/m² mayor.

La otra opción, por supuesto, es crear una nueva variable de refresco cuya unidad sea 12 onzas en lugar de 1 onza, pero la multiplicidad de la pendiente es una operación matemática simple y es mucho más eficiente. También en este caso debe haber tenido algunas expectativas, basadas en el análisis exploratorio de datos que hizo, sobre la magnitud de la relación entre el consumo de refrescos no dietéticos y el IMC, por lo que debe determinar si su interpretación de la magnitud de la relación coincide con sus expectativas. Si no es así, tendrá que determinar si sus expectativas eran incorrectas o si su interpretación era incorrecta y actuar en consecuencia para hacer coincidir las expectativas y el resultado de su interpretación.

Otra consideración importante sobre la magnitud de la relación es si es significativa. Por ejemplo, una
El aumento de 0,01 en el IMC por cada 20 onzas adicionales consumidas al día probablemente no sea especialmente significativo, ya que una gran cantidad de refresco se asocia a un aumento muy pequeño del IMC. Por otro lado, si hubiera un aumento de 0,28 kg/m²

por cada aumento de 1 onza en el consumo de refrescos, esto sería de hecho bastante significativo. Como se sabe que el IMC oscila generalmente entre la adolescencia y los 30 años, un cambio de 0,01 kg/m^2 es pequeño, pero un cambio de 0,28 kg/m^2 podría ser significativo.

En el contexto de los volúmenes de refresco que la gente puede consumir, un aumento de 0,01 kg/m^2 por cada 20 onzas de refresco es pequeño, ya que la gente (con suerte) no bebe 10 porciones de 20 onzas al día, que es lo que alguien tendría que beber para observar incluso un aumento de 0,1 kg/m^2 en el IMC. Por otro lado, un aumento de 0,28 kg/m^2 en el IMC por cada onza adicional de refresco se acumularía rápidamente en el caso de las personas que consumieran un refresco no dietético adicional de 20 onzas al día, lo que equivaldría a un aumento esperado del IMC de 5,6 kg/m^2 . Una parte clave de la interpretación de la magnitud del resultado, por tanto, es entender cómo se compara la magnitud del resultado con lo que se sabe sobre este tipo de información en la población que interesa.

Incertidumbre

Ahora que ya sabe lo que dice el modelo sobre la dirección y la magnitud de la relación entre el consumo de refrescos no dietéticos y el IMC, el siguiente paso es considerar cuál es el grado de **incertidumbre** de su respuesta. Recuerde que su modelo se ha construido para ajustarse a los datos recogidos de una *muestra* de la población general y que está utilizando este modelo para entender cómo el consumo de refrescos no dietéticos está relacionado con el IMC en la población *general* de adultos en los EE.UU.

Volvamos a nuestro ejemplo de los refrescos y el IMC, que sí implica utilizar los resultados que se obtienen en la muestra para hacer inferencias sobre cuál es la verdadera relación entre los refrescos y el IMC en

la población general de adultos en los Estados Unidos. Imaginemos que el resultado de su análisis de los datos de la muestra indica que, *dentro de su muestra*, las personas que beben una onza adicional de refresco no dietético al día tienen un IMC que es 0,28 kg/m^2 mayor que los que beben una onza menos al día. Sin embargo, ¿cómo sabe si este resultado es simplemente el "ruido" del muestreo aleatorio o si es una aproximación a la verdadera relación entre la población general?

Para evaluar si el resultado de la muestra es simplemente "ruido" aleatorio, utilizamos medidas de incertidumbre. Aunque algunos podrían esperar que todas las muestras aleatorias sirvan como excelentes sustitutos de la población general, esto no es cierto. Para ilustrar esta idea con un ejemplo sencillo, imaginemos que la prevalencia de mujeres en la población adulta de EE.UU. es del 51% y que tomamos una muestra aleatoria de 100 adultos. Esta muestra puede tener un 45% de mujeres. Imagínese que extrae una nueva muestra de 100 adultos y su muestra tiene un 53% de mujeres. Podrías extraer muchas muestras como ésta e incluso extraer muestras con un 35% o un 70% de mujeres. La probabilidad de extraer una muestra con una prevalencia de mujeres tan diferente de la prevalencia global de mujeres en la población es muy pequeña, mientras que la probabilidad de extraer una muestra que tenga cerca del 51% de mujeres es mucho mayor.

Este concepto -la **probabilidad de que la muestra refleje la respuesta de la población total varía en función de lo cerca (o lejos) que esté el resultado de la muestra del resultado real de la población total-** es la base del concepto de incertidumbre. Como no sabemos cuál es la respuesta de la población total (¡por eso hacemos el análisis en primer lugar!), es imposible expresar la incertidumbre en términos de la probabilidad de que el resultado de la muestra refleje la población total. Así que hay otros enfoques para medir la incertidumbre

que se basan en este concepto general, y a continuación analizaremos dos enfoques comunes.

Una herramienta que proporciona una medida más continua de la falta de certeza es el intervalo de confianza. Un intervalo de confianza es un rango de valores que contiene el resultado de la muestra y que tiene cierto grado de confianza en que también contiene el verdadero resultado de la población total. La mayoría de los programas informáticos de modelado estadístico proporcionan intervalos de confianza del 95%, de modo que si el IC del 95% para la estimación de la muestra de 0,28 kg/m² de arriba es de 0,15-0,42 kg/m² , la interpretación aproximada es que se puede tener un 95% de confianza en que el resultado verdadero para la población general está entre 0,15 y
0,42 kg/m² .

> Una definición más precisa del intervalo de confianza del 95% sería que, en muestras repetidas, si realizáramos este experimento muchas veces (cada vez recogiendo un conjunto de datos del mismo tamaño), un intervalo de confianza construido de esta manera cubriría la verdad el 95% de las veces. Es importante tener en cuenta que, como indica la definición, el intervalo de confianza se construye a partir de los datos, el *propio intervalo es aleatorio*. Por lo tanto, si recogiéramos nuevos datos, el intervalo que construiríamos sería ligeramente diferente. Sin embargo, la verdad, es decir, el valor poblacional del parámetro, siempre será el mismo.

Otra herramienta para medir la incertidumbre es, por supuesto, el valor p, que simplemente es la probabilidad de obtener el resultado de la muestra de 0,28 kg/m² (o más extremo) cuando la verdadera relación entre el consumo de refrescos no dietéticos y el IMC en el conjunto de la población es 0. Aunque el valor p es una medida continua de incertidumbre, mucha gente considera

un valor p de <0,05, que indica que hay menos de un 5% de probabilidad de observar el resultado de la muestra (o un resultado más extremo) cuando no hay relación en la población general, como "estadísticamente significativo". Este punto de corte es ar- bitrario y nos dice muy poco sobre el *grado* de incertidumbre o sobre dónde se encuentra la verdadera respuesta para la población global. Centrarse principalmente en el valor p es un enfoque arriesgado para interpretar la incertidumbre, ya que puede llevar a ignorar información más importante necesaria para una interpretación reflexiva y precisa de los resultados.

El IC es más útil que el valor p, porque da un rango, que proporciona una estimación cuantitativa sobre lo que es probable que sea el resultado global real de la población, y también proporciona una forma de expresar lo seguro que es que el rango contiene el resultado global de la población.

Veamos cómo se utilizaría el valor p frente al IC del 95% para interpretar la incertidumbre sobre el resultado del análisis del IMC de los refrescos. Digamos que nuestro resultado fue que el IMC era 0,28 kg/m^2 más alto de media entre nuestra muestra que bebía una onza más de refresco no dietético al día y que el valor p asociado a este resultado era 0,03. Utilizando el valor p como herramienta para medir la incertidumbre y fijando un umbral de significación estadística en 0,05, in- terpretaríamos la incertidumbre de la siguiente manera: hay menos de un 5% de posibilidades de que obtengamos este resultado (0,28) o algo más extremo si el verdadero valor poblacional fuera 0 (o en otras palabras, que realmente no hubiera una asociación entre el consumo de refrescos y el IMC en la población general).

Ahora hagamos el mismo ejercicio con el IC del 95%. El IC del 95% para este análisis es de 0,15-0,42. Utilizando el IC como herramienta para interpretar la incertidumbre, podríamos decir que tenemos un 95% de confianza en que la verdadera relación entre el consumo de refrescos y el IMC en la población adulta de EE.UU. es

entre 0,15 y 0,42 kg/m^2 de aumento del IMC de media por cada onza adicional de refresco no dietético que se consuma. El uso de este último enfoque nos dice algo sobre la gama de posibles efectos de los refrescos en el IMC y también nos dice que es muy poco probable que los refrescos no tengan ninguna relación con el IMC en la población general de adultos en los Estados Unidos. El uso del valor p como medida de incertidumbre, por otro lado, implica que sólo tenemos dos opciones en términos de interpretación del resultado: o bien hay una buena cantidad de incertidumbre al respecto, por lo que debemos concluir que no hay relación entre el consumo de refrescos y el IMC, o bien hay muy poca incertidumbre sobre el resultado, por lo que debemos concluir que hay una relación entre el consumo de refrescos y el IMC. El uso del valor p nos limita de una manera que no refleja el proceso de sopesar la fuerza de la evidencia a favor (o en contra) de una hipótesis.

Otro punto sobre la incertidumbre es que hemos discutido la evaluación de la incertidumbre a través de los enfoques estadísticos más clásicos, que se basan en el paradigma frecuentista, que es el enfoque más común. El marco bayesiano es un enfoque alternativo en el que se actualizan las creencias previas en función de las pruebas proporcionadas por el análisis. En la práctica, el enfoque frecuencial del que hablamos anteriormente es el más utilizado y, en el mundo real, rara vez conduce a conclusiones diferentes de las que se obtienen utilizando un enfoque bayesiano.

Una advertencia importante es que a veces no es necesario evaluar la falta de certeza porque algunos tipos de análisis no están destinados a hacer inferencias sobre una población general más amplia. Si, por ejemplo, quiere entender la relación entre la edad y los dólares gastados al mes en los productos de su empresa, puede tener todos los datos sobre la población completa o "general" que le interesa, que son los clientes de su empresa. En este caso, usted no

tiene que basarse en una muestra, porque su empresa recoge datos sobre la edad y las compras de TODOS sus clientes. En este caso, no tendría que considerar la incertidumbre de que su resultado refleje la verdad para la población global, porque el resultado de su análisis **es la verdad** para su población global.

Elabore una interpretación global teniendo en cuenta la totalidad de sus análisis y las información

Ahora que ha dedicado una buena cantidad de esfuerzo a inter- pretar los resultados de su modelo primario, el siguiente paso es desarrollar una interpretación global de sus resultados considerando tanto la totalidad de sus análisis como la información externa a los mismos. La interpretación de los resultados de su modelo primario sirve para establecer la expectativa de su interpretación general cuando considere todos sus análisis. Siguiendo con el ejemplo de los refrescos y el IMC, supongamos que su interpretación del modelo primario es que el IMC es 0,28 kg/m^2 más alto de media entre los adultos de EE.UU. que consumen de media una onza más de refresco al día. Recuerde que este modelo primario se construyó después de recopilar información a través de análisis exploratorios y que usted puede haber refinado este modelo cuando estaba pasando por el proceso de interpretación de sus resultados mediante la evaluación de la direccionalidad, la magnitud y la incertidumbre de los resultados del modelo.

Tal y como se ha comentado en el capítulo de modelización formal, no existe un único modelo que por sí solo proporcione **la** respuesta a su pregunta. En su lugar, hay modelos adicionales que sirven para cuestionar el resultado obtenido en el modelo primario. Un tipo común de modelo secundario es el que se construye para determinar la sensibilidad de los resultados en su

modelo primario son a los cambios en los datos. Un ejemplo clásico es la eliminación de los valores atípicos para evaluar el grado en que cambia el resultado del modelo primario. Si los resultados del modelo primario estuvieran impulsados en gran medida por un puñado de, por ejemplo, consumidores de refrescos muy elevados, este hallazgo sugeriría que puede no haber una relación lineal entre el consumo de refrescos y el IMC y que, en cambio, el consumo de refrescos sólo puede influir en el IMC entre aquellos que tienen un consumo muy elevado de refrescos. Este hallazgo debería llevar a una revisión de su modelo principal.

Un segundo ejemplo es la evaluación del efecto de los posibles factores de confusión en los resultados del modelo primario. Aunque el modelo primario ya debería contener los principales factores de confusión, normalmente hay otros posibles factores de confusión que deben evaluarse. En el ejemplo de los refrescos y el IMC, puede construir un modelo secundario que incluya los ingresos porque se da cuenta de que es posible que la relación que obtiene en su modelo primario pueda explicarse completamente por el estatus socioeconómico: las personas con un estatus socioeconómico más alto pueden beber menos refrescos no dietéticos y también tener un IMC más bajo, pero no es porque beban menos refrescos. En cambio, es algún otro factor asociado al estatus socioeconómico el que tiene el efecto sobre el IMC. Así que se puede ejecutar un modelo secundario en el que se añadan los ingresos al modelo primario para determinar si este es el caso. Aunque hay otros ejemplos de usos de modelos secundarios, estos son dos ejemplos comunes.

Entonces, ¿cómo se interpreta la forma en que estos resultados secundarios del modelo afectan al resultado principal? Puede recurrir al paradigma de: direccionalidad, magnitud e incertidumbre. Cuando se añadieron los ingresos al modelo refresco-IMC, ¿cambió la dirección de la relación estimada entre el refresco y el IMC del modelo primario, ya sea a una asociación negativa o a ninguna asociación? Si lo hizo, eso

sería un cambio dramático y sugeriría que o bien algo no está bien con sus datos (como con la variable de ingresos) o que la asociación entre el consumo de refrescos y el IMC se explica totalmente por los ingresos.

Supongamos que la adición de los ingresos no cambió la di- reccionalidad y supongamos que cambió la magnitud de modo que la estimación del modelo primario de 0,28 kg/m^2 disminuyó a 0,12kg/m^2 . La magnitud de la relación entre los refrescos y el IMC se redujo en un 57%, por lo que esto se interpretaría como que los ingresos explican algo más de la mitad, pero no toda, la relación entre el consumo de refrescos y el IMC.

Ahora se pasa a la incertidumbre. El IC del 95% para la estimación con el modelo que incluye los ingresos es de 0,01-0,23, por lo que podemos estar seguros al 95% de que la verdadera relación entre los refrescos y el IMC en la población adulta de EE.UU., independientemente de los ingresos, se encuentra en algún lugar de este rango. ¿Y si el IC del 95% para la estimación fuera -0,02-0,26, pero la estimación siguiera siendo de 0,12 kg/m^2 ? Aunque el IC incluya ahora 0, el resultado del modelo primario, 0,12, no cambió, lo que indica que los ingresos no parecen explicar nada de la asociación entre el consumo de refrescos y el IMC, pero que sí aumentó la incertidumbre del resultado. Una de las razones por las que la adición de los ingresos al modelo podría haber aumentado la incertidumbre es que a algunas personas de la muestra les faltaban datos sobre los ingresos, por lo que el tamaño de la muestra se redujo. La comprobación de sus n le ayudará a determinar si éste es el caso.

También es importante considerar los resultados generales en el contexto de la información externa. La información externa es el conocimiento general que usted o los miembros de su equipo tienen sobre el tema, los resultados de análisis similares y la información sobre la población objetivo. Un ejemplo que se ha comentado anteriormente es que tener una idea de cuáles son los volúmenes típicos y plausibles de consumo de refrescos entre los adultos en

en Estados Unidos es útil para entender si la magnitud del efecto del consumo de refrescos en el IMC es significativa. También puede ser útil saber qué porcentaje de la población adulta de EE.UU. bebe refrescos no dietéticos y la prevalencia de la obesidad para entender el tamaño de la población para la que sus resultados podrían ser pertinentes.

Un ejemplo interesante de la importancia de pensar en el tamaño de la población que puede verse afectada es la contaminación atmosférica. En el caso de las asociaciones entre la contaminación del aire exterior y los resultados sanitarios críticos, como los eventos cardiovasculares (ictus, infarto de miocardio), la magnitud del efecto es pequeña, pero dado que la contaminación del aire afecta a cientos de millones de personas en EE.UU., el número de eventos cardiovasculares atribuibles a la contaminación es bastante elevado.

Además, es probable que usted no sea la primera persona que intenta responder a esta pregunta o a otras relacionadas. Otros pueden haber realizado un análisis para responder a la pregunta en otra población (adolescentes, por ejemplo) o para responder a una pregunta relacionada, pero diferente, como por ejemplo "¿cuál es la relación entre el consumo de refrescos no dietéticos y los niveles de azúcar en sangre?" Comprender cómo encajan sus resultados en el contexto del conjunto de conocimientos sobre el tema le ayuda a usted y a otros a evaluar si existe una historia o un patrón general que emerge de todas las fuentes de conocimiento y que apunta a que el consumo de refrescos no dietéticos está relacionado con el nivel elevado de azúcar en sangre, la resistencia a la insulina, el IMC y la diabetes de tipo 2. Por otro lado, si los resultados de su análisis difieren de la base de conocimientos externa, eso también es importante. Aunque la mayoría de las veces, cuando los resultados son tan sorprendentemente diferentes de los conocimientos externos, hay una explicación, como un error o diferencias en los métodos de recogida de datos o en la población estudiada, a veces un hallazgo claramente diferente es una

visión verdaderamente novedosa.

Implicaciones

Ahora que ha interpretado los resultados y tiene las conclusiones en la mano, querrá pensar en las implicaciones de sus conclusiones. Al fin y al cabo, el objetivo de un análisis suele ser informar de una decisión o emprender una acción. A veces las implicaciones son sencillas, pero otras veces hay que pensar en ellas. Un ejemplo de una implicación directa es si se realiza un análisis para determinar si la compra de anuncios aumenta las ventas y, en caso afirmativo, si la inversión en anuncios produce un beneficio neto. Puede que descubra que hubo un beneficio neto o no, y si hubo un beneficio neto, esta conclusión apoyaría la continuación de los anuncios.

Un ejemplo más complicado es el de los refrescos y el IMC que hemos utilizado a lo largo de este capítulo. Si el consumo de refrescos se asocia a un mayor IMC, con una porción adicional de 20 onzas al día asociada a un IMC 0,28 kg/m^2 mayor, este hallazgo implicaría que si se pudiera reducir el consumo de refrescos, se podría reducir el IMC medio de la población en general. Sin embargo, dado que su análisis no fue causal y sólo demostró una asociación, es posible que desee realizar un estudio en el que asigne aleatoriamente a las personas a sustituir uno de los refrescos de 20 onzas que beben cada día por un refresco dietético o a no sustituir su refresco no dietético. Sin embargo, en un entorno de salud pública, su equipo puede decidir que esta asociación es evidencia suficiente para lanzar una campaña de salud pública para reducir el consumo de refrescos, y que no necesita datos adicionales de un ensayo clínico. En su lugar, puede planificar el seguimiento del IMC de la población durante y después de la campaña de salud pública como medio para estimar el efecto en la salud pública de la reducción del consumo de refrescos no dietéticos. La conclusión es que la acción resultante de las implicaciones a menudo depende de la misión de la organización que solicitó el análisis.

10. Comunicación

La comunicación es fundamental para un buen análisis de datos. Lo que pretendemos abordar en este capítulo es el papel de la comunicación rutinaria en el proceso de análisis de datos y en la difusión de los resultados finales en un entorno más formal, a menudo ante un público externo más amplio. Hay muchos libros buenos que abordan el "cómo" de las presentaciones formales, ya sea en forma de charla o de artículo escrito, como un libro blanco o un artículo científico. En este capítulo, sin embargo, nos centraremos en:

1. Cómo utilizar la comunicación rutinaria como una de las herramientas necesarias para realizar un buen análisis de datos.
2. Cómo transmitir los puntos clave de su análisis de datos cuando se comunica de manera informal y formal.

La comunicación es tanto una de las herramientas del análisis de datos como el producto final del mismo: no tiene sentido hacer un análisis de datos si no se va a comunicar el proceso y los resultados a un público. Un buen analista de datos se comunica de manera informal en múltiples ocasiones durante el proceso de análisis de datos y también reflexiona cuidadosamente sobre la comunicación de los resultados finales para que el análisis sea lo más útil e informativo posible para el público más amplio al que va dirigido.

10.1 Comunicación rutinaria

El objetivo principal de la comunicación rutinaria es la recopilación de datos, que forma parte del proceso epicicloidal de cada núcleo

actividad. La recopilación de datos se lleva a cabo mediante la comunicación de los resultados, y las respuestas que reciba de su público deben servir de base para los siguientes pasos del análisis de datos. Los tipos de respuestas que reciba incluyen no sólo las respuestas a preguntas específicas, sino también los comentarios y las preguntas que su audiencia tenga en respuesta a su informe (ya sea escrito u oral). La forma que adopte la comunicación rutinaria dependerá del objetivo de la misma. Si su objetivo, por ejemplo, es aclarar cómo se codifica una variable porque al explorar el conjunto de datos parece ser una variable ordinal, pero usted había entendido que era una variable continua, su comunicación es breve y directa.

Si, por el contrario, algunos de los resultados de su análisis exploratorio de datos no son los que esperaba, su comunicación puede adoptar la forma de una pequeña reunión informal que incluya la exhibición de tablas y/o figuras pertinentes a su problema. Un tercer tipo de comunicación informal es aquella en la que no se tienen preguntas específicas que hacer a la audiencia, sino que se busca una opinión sobre el proceso de análisis de datos y/o los resultados para ayudar a perfeccionar el proceso y/o informar sobre los próximos pasos.

En resumen, hay tres tipos principales de comunicación informal y se clasifican en función de los objetivos que se persiguen con la comunicación: (1) responder a una pregunta muy concreta, que suele ser una cuestión técnica o una pregunta destinada a recopilar un dato, (2) ayudar a resolver algunos resultados que son desconcertantes o no son exactamente lo que se esperaba, y
(3) para obtener impresiones y comentarios generales como medio de identificar cuestiones que no se le habían ocurrido para poder perfeccionar su análisis de datos.

Centrarse en unos pocos conceptos básicos le ayudará a alcanzar sus objetivos cuando planifique la comunicación rutinaria. Estos conceptos son:

1. **Audiencia**: Conozca a su público y, cuando tenga el control sobre quién es el público, seleccione la audiencia adecuada para el tipo de respuesta que está buscando.
2. **Contenido**: Sé centrado y conciso, pero proporciona suficiente información para que la audiencia entienda la información que presentas y la(s) pregunta(s) que planteas.
3. **Estilo**: Evite la jerga. A no ser que se trate de un tema muy técnico y se dirija a un público muy técnico, es mejor utilizar un lenguaje y unas cifras y tablas que puedan ser entendidas por un público más general.
4. **Actitud**: Ten una actitud abierta y colaboradora para que estés dispuesto a participar plenamente en un diálogo y para que tu audiencia reciba el mensaje de que tu objetivo no es "defender" tu pregunta o tu trabajo, sino obtener su opinión para que puedas hacer tu mejor trabajo.

10.2 El público

Para muchos tipos de comunicación rutinaria, podrá seleccionar su audiencia, pero en algunos casos, como cuando presenta un informe provisional a su jefe o a su equipo, la audiencia puede estar predeterminada. Su audiencia puede estar compuesta por otros analistas de datos, la(s) persona(s) que inició(n) la pregunta, su jefe y/o otros gerentes o miembros del equipo ejecutivo, personas que no son analistas de datos pero que son expertos en el contenido, y/o alguien que represente al público en general.

Para el primer tipo de comunicación rutinaria, en la que se busca principalmente un conocimiento fáctico o una aclaración sobre el conjunto de datos o la información relacionada, se debe seleccionar una persona (o personas) que tenga el conocimiento fáctico para responder a la

y que respondan a las preguntas es lo más apropiado. Para una pregunta sobre cómo se recogieron los datos de una variable en el conjunto de datos, puede dirigirse a una persona que haya recogido los datos o a una persona que haya trabajado antes con el conjunto de datos o que haya sido responsable de la compilación de los datos. Si la pregunta es sobre el comando que hay que utilizar en un lenguaje de programación estadística para ejecutar un determinado tipo de prueba estadística, esta información suele encontrarse fácilmente mediante una búsqueda en Internet. Pero si esto falla, sería apropiado consultar a una persona que utilice el lenguaje de programación en cuestión.

Para el segundo tipo de comunicación rutinaria, en la que se tienen algunos resultados y no se está seguro de si son los que se esperaban o no son los que se esperaban, lo más probable es que sea de gran ayuda si se cuenta con la participación de más de una persona que represente una variedad de perspectivas. Las reuniones más productivas y útiles suelen incluir a personas con experiencia en el análisis de datos y en el área de contenidos. Como regla general, cuantos más tipos de partes interesadas se comuniquen durante el proyecto de análisis de datos, mejor será el producto final. Por ejemplo, si sólo te comunicas con otros analistas de datos, puedes pasar por alto algunos aspectos importantes de tu análisis de datos que habrías descubierto si te hubieras comunicado con tu jefe, con expertos en contenidos o con otras personas.

Para el tercer tipo de comunicación rutinaria, que suele producirse cuando se ha llegado a un punto natural para hacer una pausa en el análisis de los datos. Aunque el momento y el lugar en el que se producen estas pausas dependen del análisis específico que se esté realizando, un lugar muy común para hacer una pausa y hacer balance es después de completar al menos un análisis exploratorio de los datos. Es importante hacer una pausa y pedir opiniones en este punto, ya que este ejercicio suele identificar análisis exploratorios adicionales que son importantes para in-

formar los siguientes pasos, como la construcción de modelos, y así evitar que se pierda tiempo y esfuerzo en la búsqueda de modelos que no son relevantes, no son apropiados, o ambas cosas. Este tipo de comunicación es más eficaz cuando adopta la forma de una reunión cara a cara, pero las videoconferencias y las conversaciones telefónicas también pueden ser efectivas. A la hora de seleccionar a su público, piense en quiénes de entre las personas de las que dispone le aportan los comentarios más útiles y qué perspectivas serán importantes para informar de los siguientes pasos de su análisis. Como mínimo, deberían estar representados tanto los expertos en análisis de datos como los expertos en contenido, pero en este tipo de reuniones también puede ser útil escuchar a personas que compartan, o al menos entiendan, la perspectiva del público objetivo más amplio para la comunicación formal de los resultados de su análisis de datos.

10.3 Contenido

El principio rector más importante es adaptar la información que se ofrece al objetivo de la comunicación. En el caso de una pregunta dirigida a obtener aclaraciones sobre la codificación de una variable, el destinatario de su comunicación no necesita conocer el objetivo general de su análisis, ni lo que ha hecho hasta ese momento, ni ver ninguna figura o tabla. Una pregunta específica y concreta del tipo "Estoy analizando el conjunto de datos sobre delincuencia que me enviaste la semana pasada y estoy mirando la variable "educación" y veo que está codificada como 0, 1 y 2, pero no veo ninguna etiqueta para esos códigos. ¿Sabe usted qué significan esos códigos para la variable "educación"?"

Para el segundo tipo de comunicación, en la que se busca una respuesta a causa de un problema desconcertante o inesperado en el análisis, se necesitará más información de fondo

necesario, pero puede que la información de fondo completa para el proyecto general no lo sea. Para ilustrar este concepto, supongamos que ha estado examinando la relación entre la altura y la función pulmonar y construye un gráfico de dispersión que sugiere que la relación no es lineal, ya que parece haber una curvatura en la relación. Aunque tiene algunas ideas sobre los enfoques para tratar las relaciones no lineales, busca apropiadamente la opinión de otros. Después de reflexionar sobre los objetivos de la comunicación, se decide por dos objetivos principales: (1) Comprender si existe un enfoque óptimo para tratar la no linealidad de la relación y, en caso afirmativo, cómo determinar qué enfoque es el mejor, y (2) Comprender más sobre la relación no lineal que observa, incluyendo si es esperada y/o conocida y si es importante captar la no linealidad en sus análisis.

Para lograr sus objetivos, tendrá que proporcionar a su audiencia algo de contexto y antecedentes, pero proporcionar antecedentes completos del proyecto de análisis de datos y revisar todos los pasos que ha dado hasta ahora es innecesario y probablemente absorba tiempo y esfuerzo que sería mejor dedicar a sus objetivos específicos. En este ejemplo, el contexto y los antecedentes adecuados podrían incluir lo siguiente:
(1) el objetivo general del análisis de datos, (2) cómo encajan la altura y la función pulmonar en el objetivo general del análisis de datos, por ejemplo, la altura puede ser un posible factor de confusión, o el principal predictor de interés, y (3) lo que ha hecho hasta ahora con respecto a la altura y la función pulmonar y lo que ha aprendido. Este último paso debería incluir alguna presentación visual de los datos, como el gráfico de dispersión antes mencionado. El contenido final de su presentación, por lo tanto, incluiría una declaración de los objetivos de la discusión, un breve resumen del proyecto de análisis de datos, cómo el problema específico que está enfrentando encaja en el proyecto general de análisis de datos, y

y, por último, los resultados pertinentes de su análisis relacionados con la altura y la función pulmonar.

Si estuviera elaborando una presentación de diapositivas, debería dedicar menos diapositivas a los antecedentes y al contexto que a la presentación de los resultados del análisis de datos sobre la estatura y la función pulmonar. Una diapositiva debería ser suficiente para el resumen del análisis de datos, y 1-2 diapositivas deberían ser suficientes para explicar el contexto del tema de la altura y la función pulmonar dentro del proyecto de análisis de datos más amplio. El grueso de la presentación no debería requerir más de 5-8 diapositivas, por lo que el tiempo total de la presentación no debería superar los 10-15 minutos. Aunque las diapositivas no son ciertamente necesarias, una herramienta visual para presentar esta información es muy útil y no debe implicar que la presentación deba ser "for- mal". En cambio, la idea es proporcionar al grupo información suficiente para generar un debate centrado en sus objetivos, lo que se consigue mejor con una presentación informal.

Estos mismos principios se aplican al tercer tipo de comunicación, con la salvedad de que es posible que no tenga objetivos concretos y que, en cambio, busque la opinión general de su público sobre su proyecto de análisis de datos. Si este es el caso, debe indicarse este objetivo más general y el resto del contenido debe incluir una declaración de la pregunta a la que pretende responder con el análisis, el objetivo o los objetivos del análisis de datos, un resumen de las características del conjunto de datos (fuente de los datos, número de observaciones, etc.), un resumen de sus análisis exploratorios, un resumen de la construcción del modelo, su interpretación de los resultados y las conclusiones. Al proporcionar los puntos clave de todo el análisis de los datos, el público podrá aportar su opinión sobre el proyecto en general y sobre cada uno de los pasos del análisis de los datos. Un debate bien planificado produce comentarios útiles y reflexivos y debe considerarse un éxito si

se le dejará armado con los refinamientos adicionales que debe hacer a su análisis de datos y la perspectiva reflexiva sobre lo que debe incluirse en la presentación más formal de sus resultados finales a una audiencia externa.

10.4 Estilo

Aunque el estilo de comunicación aumenta en formalidad desde el primer al tercer tipo de comunicación rutinaria, todas estas comunicaciones deben ser en gran medida informales y, salvo quizás la comunicación centrada en un pequeño asunto técnico, debe evitarse la jerga. Dado que el objetivo principal de la comunicación rutinaria es obtener retroalimentación, su estilo de comunicación debe fomentar el debate. Algunos enfoques para fomentar el debate incluyen declarar por adelantado que le gustaría que la mayor parte de la reunión incluyera un debate activo y que agradece las preguntas durante su presentación en lugar de pedir a la audiencia que las retenga hasta el final de su presentación. Si un miembro del público hace un comentario, preguntar qué piensan los demás también fomentará el debate. En esencia, para obtener la mejor retroalimentación, usted quiere escuchar lo que los miembros de su audiencia piensan, y esto se logra más probablemente estableciendo un tono informal y fomentando activamente la discusión.

10.5 Actitud

Una actitud defensiva o desagradable puede sabotear todo el trabajo que ha realizado al seleccionar cuidadosamente el público, al identificar cuidadosamente sus objetivos y preparar su conferencia, y al afirmar que está buscando un debate. El público se mostrará reacio a ofrecer comentarios constructivos si

sensación de que sus comentarios no serán bien recibidos y saldrás de la reunión sin haber alcanzado tus objetivos, y mal preparado para hacer cualquier refinamiento o adición a tu análisis de datos. Y cuando llegue el momento de hacer una presentación formal ante un público externo, no estarás bien preparado y no podrás presentar tu mejor trabajo. Para evitar este escollo, cultiva deliberadamente una actitud receptiva y positiva antes de la comunicación, dejando a un lado tu ego y tus inseguridades. Si consigues hacerlo con éxito, te servirá de mucho. De hecho, ambos conocemos a personas que han tenido carreras muy exitosas basadas en gran medida en su actitud positiva y acogedora hacia la retroalimentación, incluida la crítica constructiva.

11. Reflexiones finales

Ahora debería disponer de un enfoque que puede aplicar a sus análisis de datos. Aunque cada conjunto de datos es un organismo único y cada análisis tiene sus propios problemas específicos a los que enfrentarse, abordar cada paso con el marco del epiciclo es útil para cualquier análisis. Mientras trabaja en el desarrollo de su pregunta, en la exploración de sus datos, en la modelización de sus datos, en la interpretación de sus resultados y en la comunicación de sus resultados, recuerde siempre establecer expectativas y luego comparar el resultado de su acción con sus expectativas. Si no coinciden, identifica si el problema está en el resultado de tu acción o en tus expectativas y corrige el problema para que coincidan. Si no puedes identificar el problema, busca la opinión de otros y, cuando hayas solucionado el problema, pasa a la siguiente acción. Este marco epicicloidal te ayudará a mantener un rumbo que acabe en una respuesta útil a tu pregunta.

Además del marco del epiciclo, también hay actividades de análisis de datos que hemos discutido a lo largo del libro. Aunque todas las actividades de análisis son importantes, si tuviéramos que identificar las más importantes para asegurar que su análisis de datos proporciona una respuesta válida, significativa e interpretable a su pregunta, incluiríamos las siguientes:

1. Piense bien en el desarrollo de su pregunta y utilícela como guía en todos los pasos del análisis.
2. Sigue el ABC:
    1. Compruebe siempre

2. Siempre hay que ser desafiante
3. Comunicar siempre

La mejor manera de que el marco del epiciclo y estas actividades se conviertan en algo natural es realizar muchos análisis de datos, por lo que le animamos a que aproveche las oportunidades de análisis de datos que se le presenten. Aunque, con la práctica, muchos de estos principios se convertirán en algo natural para usted, hemos comprobado que repasar estos principios nos ha ayudado a resolver una serie de problemas a los que nos hemos enfrentado en nuestros propios análisis. Esperamos, por tanto, que el libro siga siendo un recurso útil después de que haya terminado de leerlo cuando se encuentre con los tropiezos que se producen en todos los análisis.

12. Sobre los autores

Roger D. Peng es profesor asociado de bioestadística en la Escuela de Salud Pública Bloomberg de la Universidad Johns Hopkins. También es cofundador de la especialización en ciencia de datos de Johns Hopkins[1]que ha matriculado a más de 1,5 millones de estudiantes, y del blog Simply Statistics[2] , donde escribe sobre estadística y ciencia de datos para el público en general. Se puede encontrar a Roger en Twitter y GitHub @rdpeng[3].

Elizabeth Matsui es profesora de Pediatría, Epidemiología y Ciencias de la Salud Ambiental en la Universidad Johns Hopkins y alergóloga/inmunóloga pediátrica en ejercicio. Dirige, junto con el Dr. Peng, un centro de gestión y análisis de datos que apoya estudios epidemiológicos y ensayos clínicos, y es cofundadora de Skybrude Consulting, LLC[4]una empresa de consultoría de ciencia de datos. Se puede encontrar a Elizabeth en Twitter @eliza68[5].

[1]http://www.coursera.org/specialization/jhudatascience/1
[2]http://simplystatistics.org/
[3]https://twitter.com/rdpeng
[4]http://skybrudeconsulting.com
[5]https://twitter.com/eliza68