

RAZONAMIENTO ESTADÍSTICO

¿Cómo separar el grano de la paja?

La estadística trata sobre datos. Éstos son números, pero no sólo son eso. Los datos son números en un contexto. El número 3,75, por ejemplo, no contiene por sí mismo ninguna información. Pero si oímos que una amiga ha dado a luz un bebé de 3,75 kilos, la felicitamos ya que es un buen peso. El contexto nos permite sacar partido de nuestros conocimientos sobre el tema de estudio y emitir juicios. Sabemos que un peso de 3,75 kilos es un peso adecuado (no es posible que un bebé pese 3,75 gramos o 3,75 toneladas). El contexto hace que el número aporte información.

La estadística utiliza datos para profundizar en un tema y sacar conclusiones. Nuestras herramientas son gráficos y cálculos. Son herramientas dirigidas por una línea de pensamiento basada en el sentido común. Empecemos nuestro estudio de la estadística repasando informalmente los principios de los razonamientos estadísticos.

Los datos iluminan

De la población estadounidense, ¿qué porcentaje crees que es de raza negra? ¿Qué porcentaje crees que es de raza blanca? Se formuló esta pregunta a una muestra de estadounidenses de raza blanca, la media de sus respuestas fue 23,8% negros y 49,9 blancos. Sin embargo, los datos del censo indican que el 11,8% de los estadounidenses son negros y el 74% blancos.¹

La raza es uno de los temas más conflictivos en EE UU. Es ilustrativo ver que los blancos creen, de forma equivocada, que son una minoría, cuando los datos

¹New York Times, 25 de marzo de 1996.

del censo demuestran que en realidad son la gran mayoría. Los datos del censo iluminan y muestran las cosas tal como son. Seguramente, si la gente conociera más los hechos y no se basara en suposiciones, cambiarían muchas actitudes.

Datos contra anécdotas

Una anécdota es una historia sorprendente que precisamente recordamos por serlo. Las anécdotas ayudan a humanizar los temas; sin embargo, pueden conducir a engaño.

Vivir cerca de líneas de alta tensión, ¿causa leucemia a los niños? El Instituto Nacional del Cáncer de EE UU invirtió 5 años y 5 millones de dólares para intentar responder a esta pregunta. El resultado: no existe ninguna relación entre la leucemia y la exposición a campos magnéticos como los que producen las líneas de alta tensión. En el artículo que se publicó en el *New England Journal of Medicine* se podía leer: “ha llegado el momento de no malgastar más recursos en este asunto”.²

Comparemos el impacto de una noticia aparecida en televisión sobre una investigación de 5 años y 5 millones de dólares, con el impacto de una entrevista a una madre con facilidad de palabra que vive cerca de una línea de alta tensión cuyo hijo tiene leucemia. Para la gente que ve la televisión la entrevista a la madre deja una huella mucho más profunda. Sin embargo, una persona que sepa estadística sabe que los datos son más fiables que las anécdotas ya que describen de forma sistemática una situación general y no se centran en ningún incidente concreto.

Ojo con las variables latentes

La gente que utiliza con frecuencia los aviones quiere llegar a la hora. Las compañías aéreas envían datos al Ministerio de Transportes sobre el número de vuelos que llegan a tiempo. He aquí cifras de un determinado mes correspondiente a dos compañías aéreas estadounidenses:

²E. W. Campion, “Editorial: power lines, cancer, and fear”, *New England Journal of Medicine*, 337, nº 1, 1997. El trabajo citado es M. S. Linet *et al.*, “Residential exposure to magnetic fields and acute lymphoblastic leukemia in children”. Lo encontré en <<http://www.nejm.org>>. Consulta también G. Taubes, “Magnetic field-cancer link: will it rest in peace?”, *Science*, 277, 1997, pág. 29.

	Vuelos puntuales	Vuelos con retraso
Alaska Airlines	3.274	501
America West	6.438	787

Puedes comprobar que el porcentaje de vuelos que llegaron con retraso es

$$\text{Alaska Airlines} \quad \frac{501}{3.775} = 13,3\%$$

$$\text{America West} \quad \frac{787}{7.225} = 10,9\%$$

Aparentemente la mejor compañía es America West.

Sin embargo, hay que tener en cuenta más información. Casi todas las relaciones entre dos variables dependen de otras latentes de entorno. Tenemos datos de dos variables: la compañía aérea y si los vuelos llegan o no a la hora. Tengamos en cuenta una tercera variable: la ciudad de procedencia del vuelo.³

	Alaska Airlines		America West	
	Vuelos puntuales	Vuelos con retraso	Vuelos puntuales	Vuelos con retraso
Los Angeles	497	62	694	117
Phoenix	221	12	4.840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1.841	305	201	61
Total	3.274	501	6.438	787

La fila "Total" muestra que los datos de la nueva tabla son los mismos que los datos de la tabla anterior. Fijémonos otra vez en los porcentajes de vuelos con retraso; los que proceden de los Angeles son

$$\text{Alaska Airlines} \quad \frac{62}{559} = 11,1\%$$

$$\text{America West} \quad \frac{117}{811} = 14,4\%$$

³A. Barnett, "How numbers can trick you", *Technology Review*, octubre 1994, págs. 38-45.

Alaska Airlines es la mejor compañía. Los porcentajes de retrasos de los vuelos que proceden de Phoenix son

$$\text{Alaska Airlines} \quad \frac{12}{233} = 5,2\%$$

$$\text{America West} \quad \frac{415}{5.255} = 7,9\%$$

Otra vez, la mejor compañía es Alaska Airlines. La figura 1 muestra que en todos los aeropuertos de procedencia considerados, Alaska Airlines es la compañía con un menor porcentaje de vuelos con retraso.

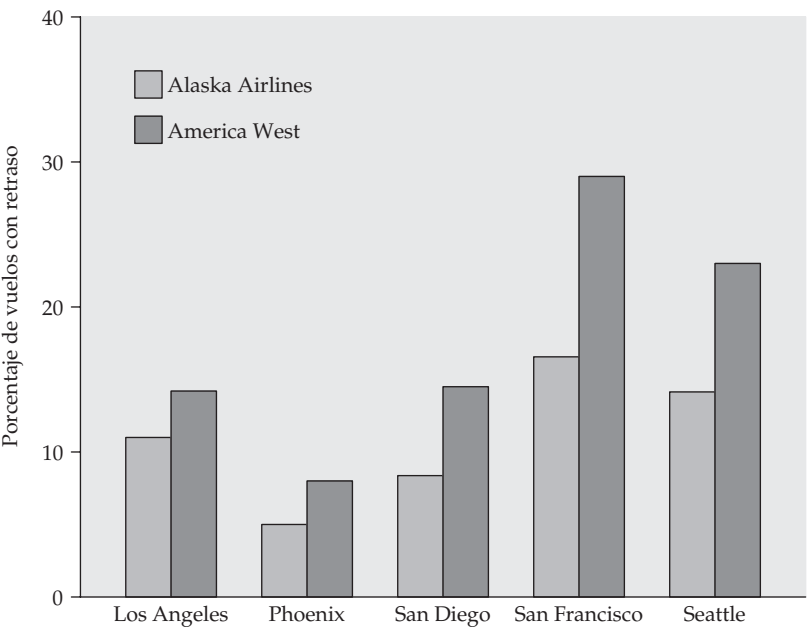


Figura 1. Comparación del porcentaje de vuelos con retraso de dos compañías en cinco aeropuertos.

¿Cómo es posible que cuando analizamos los datos aeropuerto por aeropuerto Alaska Airlines sea la mejor compañía, pero que cuando combinamos todos los datos lo sea American West? Fíjate en los datos: America West es la compañía con más vuelos procedentes de la soleada Phoenix, aeropuerto en el que se producen pocos retrasos. Sin embargo, Alaska Airlines es la compañía con más vuelos procedentes de Seattle, aeropuerto en el que la niebla obliga a retrasar numerosos

vuelos. La ciudad de procedencia influye sobre la posibilidad de sufrir un retraso en un vuelo. En consecuencia, tener en cuenta la ciudad de procedencia cambia completamente nuestras conclusiones. Vale la pena repetir el mensaje: casi todas las relaciones entre dos variables dependen de variables latentes de entorno.

Es importante saber cómo se obtuvieron los datos

En una ocasión, la periodista Ann Landers preguntó a sus lectores: “Si pudieras retroceder en el tiempo, ¿volverías a tener hijos?”. Unas semanas más tarde apareció en grandes titulares: “EL 70% DE LOS PADRES DICE QUE NO VALE LA PENA TENER HIJOS”. En realidad, lo que dijo el 70% de los casi 10.000 padres que escribieron a Ann Landers fue que si pudieran escoger de nuevo, no tendrían hijos. ¿Realmente crees que el 70% de los padres lamenta haber tenido hijos?

No deberías creerlo. Las personas que respondieron estaban especialmente motivadas, tanto que se tomaron la molestia de escribir a Ann Landers. Sus cartas indicaban que muchos de ellos estaban furiosos con sus hijos. Esta personas no representan a todos los padres. No es sorprendente que, unos meses más tarde, una encuesta de opinión diseñada estadísticamente sobre el mismo tema, que daba a todos los padres la misma posibilidad de ser escogidos, hallase que el 91% de los padres volvería a tener hijos. Conclusión: si no vas con cuidado cuando obtienes los datos, puedes anunciar que un 70% dijo “No”, cuando en realidad un 90% opinaba que “Sí”.

La siempre presente variación

Dice el director de ventas: “Felicidades, vamos a brindar con cava del Penedés: este mes nuestras ventas han subido un 2% con relación a las del mes pasado. El mes pasado tuve que despedir a la mitad de nuestros vendedores, ya que las ventas bajaron un 1%”. Aunque este comentario es algo exagerado, no deja de ser cierto que muchos gerentes reaccionan de forma desmesurada ante pequeñas variaciones de indicadores importantes de la marcha de la empresa. He aquí los comentarios del director de una de las principales empresas de investigación de mercados de EE UU en relación con su experiencia:

Demasiada gente de negocios da la misma importancia a todos los datos que aparecen en un informe. Aceptan los datos como si fueran verdades absolutas. Tienen dificultades

*para tratar con el concepto de probabilidad. No ven los datos como representantes de un intervalo de valores que describen nuestro conocimiento sobre unas determinadas circunstancias.*⁴

Indicadores empresariales como las ventas y los precios varían de mes a mes por motivos tan variados como el tiempo, las dificultades financieras de los clientes o los inevitables errores en la obtención de datos. Los responsables empresariales tienen que ser capaces de identificar las variaciones realmente importantes. Las herramientas estadísticas pueden ayudar. A veces, es suficiente con representar gráficamente los datos. La figura 2 muestra el precio medio mensual del litro de gasolina en EE UU durante 10 años. Más allá de la variación mes a mes como la subida de precios en verano, el pico debido a la guerra del Golfo de 1990, o la bajada de la primavera de 1998, se puede observar una tendencia creciente gradual.⁵

La variación está siempre presente; los individuos varían, las mediciones repetidas de un mismo individuo varían; casi todo varía a lo largo del tiempo. Una de las razones de aprender estadística es que nos ayuda a tratar con la variación.

No hay verdades absolutas

A partir de una cierta edad, la mayoría de mujeres se hacen mamografías de forma regular para la detección del cáncer de mama. Las mamografías, ¿realmente reducen el riesgo de muerte? Para hallar respuestas, los médicos confían en “experimentos clínicos aleatorizados” para comparar distintos métodos de detección del cáncer de mama. Más adelante veremos que los datos procedentes de experimentos comparativos aleatorizados ayudan a dar respuestas. Las conclusiones de 13 de estos experimentos es que las mamografías reducen el riesgo de muerte entre 50 y 64 años en un 26%.⁶

⁴A. C. Nielsen, Jr., “Statistics in marketing”, en *Making Statistics More Effective in Schools of Business*, Graduate School of Business, University of Chicago, 1986.

⁵Los datos de la figura 2 se basan en un componente del Índice de Precios al Consumo, elaborado por el Bureau of Labor Statistics. Se pueden consultar estos datos en <http://stats.bls.gov:80/datahome.htm>. Para pasar los números índice a dólares, he utilizado la media del precio nacional de 1,131 dólares por galón de enero de 1998. Información obtenida de la Energy Information Agency, página web <http://www.eia.doe.gov>.

⁶H. C. Sox, “Editorial: benefit and harm associated with screening for breast cancer”, *New England Journal of Medicine*, 338, n° 16, 1998.

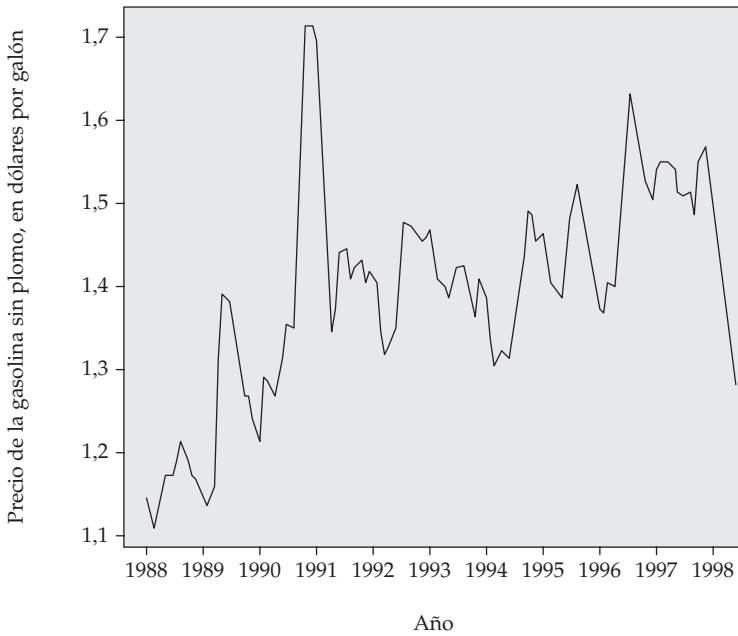


Figura 2. La variación está siempre presente: precio de la gasolina sin plomo en un surtidor de EE UU desde 1988 hasta mediados de 1998.

Como media, las mujeres que se hacen mamografías de forma regular tienen menos probabilidades de morir de cáncer de mama. Sin embargo, debido a que la variación está siempre presente, los resultados no son los mismos para todas las mujeres. Algunas mujeres que se hacen mamografías cada año mueren de cáncer de mama y otras que nunca se las han hecho pueden vivir hasta los 100 años y morir, por ejemplo, en un accidente de tráfico. ¿Podemos estar seguros de que como media las mamografías reducen el cáncer de mama? No podemos estar seguros. *Como la variación está siempre presente, no existen verdades absolutas.*

La estadística nos proporciona el lenguaje necesario para tratar con la incertidumbre, es el lenguaje utilizado por la gente que conoce estadística. En el caso de las mamografías, los médicos utilizan este lenguaje para decirnos que “las mamografías reducen el riesgo de morir de cáncer de mama en un 26 por ciento (con una confianza del 95%, la reducción se sitúa entre el 17 y el 34%)”. Este 26% son, según Arthur Nielsen, “valores que representan un intervalo que describe nuestro conocimiento sobre unas determinadas circunstancias”. El intervalo va del 17 al 34% y tenemos una confianza del 95% de que el verdadero valor se halla en este

intervalo. Pronto aprenderemos a comprender este lenguaje. No podemos evitar la variación y la incertidumbre. Aprender estadística nos permite desenvolvernó mejor ante esta realidad.

Tú y el razonamiento estadístico

Lo que encontrarás en este libro

El objetivo de este libro es que conozcas los conceptos y aprendas a utilizar la estadística aplicada. Dividiremos la estadística aplicada en tres partes que concuerdan con nuestra breve introducción a los razonamientos estadísticos.

1. **El análisis de datos** se ocupa de los métodos y las ideas necesarias para organizar y describir datos utilizando gráficos y resúmenes numéricos. Solamente los datos bien organizados iluminan. Sólo una exploración sistemática de los datos permite combatir las variables latentes. Los capítulos 1 y 2 tratan sobre el análisis de datos.
2. **La obtención de datos** proporciona métodos para obtener datos que permiten dar respuestas claras a preguntas concretas. Es muy importante saber cómo se obtuvieron los datos —los conceptos básicos sobre cómo obtener muestras y diseñar experimentos son quizás las ideas estadísticas que han tenido una mayor influencia—. Estos conceptos son el tema del capítulo 3.
3. **La inferencia estadística** va más allá de los datos disponibles y obtiene conclusiones sobre un universo más amplio, teniendo en cuenta la omnipresencia de la variabilidad y la incertidumbre de las conclusiones. Para describir la variabilidad y la incertidumbre, la inferencia estadística utiliza el lenguaje de la probabilidad que introducimos en el capítulo 4 y desarrollamos en el capítulo 5. Como estamos más interesados en la práctica que en la teoría, no es necesario un amplio conocimiento sobre probabilidad. El capítulo 6 discute los razonamientos utilizados en inferencia estadística, y los capítulos 7 y 8 presentan la inferencia tal como se utiliza en la práctica en algunas situaciones sencillas. Los capítulos 9, 10 y 11 aportan una breve introducción a la inferencia en algunas situaciones más complejas. El capítulo 12 introduce las pruebas no paramétricas.

Como los datos son números con un contexto, trabajar en estadística significa algo más que la simple manipulación de números. Este libro está lleno de datos

y cada conjunto de ellos va acompañado de un breve resumen del contexto que nos ayuda a comprender lo que dicen. En este texto, los ejemplos y los ejercicios tratan de expresar brevemente la información obtenida a partir de los datos. En la práctica, sabrías mucho más sobre el contexto en el que se obtuvieron los datos y sobre las preguntas a las que esperas que respondan los datos. Ningún libro de texto puede ser completamente realista. De todas formas, es más importante adquirir el hábito de preguntar “¿qué me dicen los datos?”, que concentrarse sólo en hacer gráficos y cálculos. Este libro trata de estimular la adquisición de buenas costumbres.

La estadística implica hacer muchos cálculos y gráficos. Este texto presenta las técnicas que necesitas. Deberías utilizar una calculadora o un ordenador para automatizar al máximo los cálculos y los gráficos. Existen muchos tipos de programas estadísticos, desde las hojas de cálculo hasta los complejos programas informáticos, idóneos para los usuarios más avanzados de la estadística. El tipo de programas estadísticos disponibles para los estudiantes varía mucho de un lugar a otro —sin embargo, las grandes ideas de la estadística no dependen del nivel de los programas de que se dispone—. A pesar de que alentamos a utilizar programas estadísticos, este libro no exige su utilización ni está ligado a ningún programa en concreto.

Este libro exige que tengas una calculadora con algunas funciones estadísticas. Concretamente, necesitas una calculadora que halle medias y desviaciones típicas, y que calcule correlaciones y rectas de regresión. Existen calculadoras con muchas funciones estadísticas, incluyendo gráficos, pero el mayor tamaño de las pantallas y la facilidad de edición de los ordenadores, los hacen más adecuados para análisis estadísticos más elaborados.

Las calculadoras y los ordenadores pueden hacer gráficos y cálculos de forma más rápida y precisa que las personas. De todas formas, como en la práctica los gráficos y los cálculos se hacen automáticamente, el activo más importante que puedes obtener a partir del estudio de la estadística es la comprensión de las grandes ideas de la estadística y la adquisición de un criterio para trabajar con datos. Las ideas y el criterio no se pueden (al menos hasta ahora) automatizar, pero te sirven de guía para indicar al ordenador lo que tiene que hacer y para interpretar los resultados obtenidos. Este libro trata de explicar las ideas más importantes de la estadística, no sólo enseña métodos estadísticos. Algunos ejemplos de grandes ideas que encontrarás en este libro (una de cada uno de los tres campos de estudio de la estadística aplicada) son “representa siempre tus datos gráficamente”, “los experimentos comparativos aleatorizados” y “la significación estadística”.

Aprendes estadística resolviendo problemas prácticos. Este libro ofrece problemas de tres niveles dispuestos de manera que te faciliten tu aprendizaje. Después de la introducción de cada idea, los problemas de “Aplica tus conocimientos” son ejercicios de aplicación directa que te permiten consolidar los conceptos a medida que se van introduciendo. Antes de seguir leyendo haz algunos de estos ejercicios. Los ejercicios al final de cada sección numerada te ayudan a integrar todas las ideas introducidas en la sección. Finalmente, los ejercicios de repaso de cada capítulo presentan una selección de ejercicios referentes a todo el capítulo. Cada nivel de ejercicios implica una mayor dificultad en cuanto a decidir qué ideas y procedimientos hay que utilizar para resolver cada ejercicio, por tanto es necesaria una mejor asimilación de contenidos. Cada capítulo termina con un “Repaso del capítulo” que contiene una lista de las habilidades que debes haber adquirido. Repasa esta lista y comprueba que efectivamente eres capaz de hacer todo lo que se cita. A continuación intenta hacer algunos ejercicios de repaso del capítulo.

De todas formas, *el principio básico del aprendizaje es la constancia*. Las principales ideas de la estadística, al igual que las principales ideas de cualquier campo de estudio importante, requieren mucho tiempo para ser descubiertas y algo más de tiempo para dominarlas. Los resultados compensarán el esfuerzo.