Actas de las Jenui, vol. 2. 2017.
Páginas: 79-86

Cuando lo estadísticamente significativo ni es estadístico ni significativo. Errores habituales al usar estadísticas.

Irene García Mosquera, Arnau Mir, José Miró-Julià
Dept. de Matemàtiques i Informàtica
Universitat de les Illes Balears

Resumen

La naturaleza de la educación, con su variabilidad e incertidumbre inherentes, obligan a que se use la estadística para analizar los datos que salen de cualquier investigación experimental. Pero el razonamiento estadístico es probabilístico, esencialmente diferente al científico o ingenieril, y esto da lugar a menudo a errores tanto en el diseño del experimento como en el análisis de los resultados. Estos errores en general provienen de considerar la estadística como un oráculo que te puede determinar con seguridad la certeza o falsedad de hipótesis. Y la estadística no permite esto. La estadística gestiona la incertidumbre, pero no puede eliminarla.

En esta ponencia mostramos algunos errores habituales en la generación e interpretación de experimentos de estadística, delimitando qué es lo que puede hacer la estadística y qué es lo que debe hacer el lector o investigador.

Abstract

The nature of education, with its inherent variability and uncertainty, force the use of statistic methods to analyze the data obtained in experimental research. But statistical reasoning is probabilistic, and threfore essentially different form scientific or engineering reasoning, and this often times produces errors both in the design of the experiment and the interpretation of the results. This errors usually occurr because the researcher considers statistics to be an oracle that allows him to determine with certainty if a hypothesis is true or false. But statistics cannot do this. Statistics manage uncertainty, but cannot eliminate it.

In this paper we show some of the common errors in the generation and interpretation of statistical experiments, marking out what is that statistics can do, and what must be done by the reader or researcher.

Palabras clave

Estadística, significación estadística, estadística bayesiana.

1. La estadística no es un oráculo

Si hacemos una pregunta, a todos nos gusta una respuesta clara: un sí o un no. Pero esto es difícil en educación: cada alumno entiende algo diferente de la misma explicación y de los mismos materiales; repetimos el mismo curso con el mismo método el mismo día mañana y tarde y los resultados se parecen poco. La investigación en educación vive en la incertidumbre y por eso usamos la estadística, que es la ciencia que nos permite gestionar y cuantificar la incertidumbre [4].

Gestionarla y cuantificarla, pero no reducirla. La estadística no permite hacer afirmaciones como «El método A da lugar a menos abandonos que el método B»; ni siquiera a establecer con certeza un rango de valores al estilo de «La media de calificaciones está entre 6,77 y 7,14». La estadística no permite dar una respuesta de «sí» o «no»: la estadística no es un oráculo.

Pero esta hambre de respuestas claras y contundentes nos hace tratar los resultados estadísticos como si fuesen oráculos. Muchos de los errores en el uso de la estadística y en la interpretación de resultados estadísticos provienen de esta ilusión de certidumbre. Creemos que la nota media es un valor representativo del conocimiento de la clase; creemos que el intervalo de confianza nos da con seguridad dónde se encuentra la proporción real de aprobados; creemos que si el p-valor es lo suficientemente pequeño, podemos asegurar que el número de trabajos entregados por alumno crece. Creemos en suma que la estadística nos permite demostrar que un nuevo método es mejor (o peor) que el antiguo. Desgraciadamente, la estadística no permite hacer nada de esto.

No es una cuestión de estudiarse mejor las definiciones y conceptos de estadística, sino pensar de una forma diferente. Kahneman y Tverski [8] basaron su exitosa carrera en demostrar una y otra vez que el lenguaje probabilístico y el razonamiento estadístico están

muy alejados de nuestra intuición. No es fácil crear o leer análisis estadísticos.

Este artículo está dirigido a los profesores que, no siendo expertos en estadística, la necesitan, ya sea para poder leer e interpretar correctamente los resultados de otros, ya sea para poder hacer análisis estadísticos en sus investigaciones. Pretendemos mostrar cómo la visión de la estadística como oráculo es la fuente de muchos de los errores típicos. Identificaremos estos errores y daremos algunos consejos que humildemente esperamos ayude a entender mejor qué puede y no puede hacer la estadística y facilitar así al investigador en docencia el diseño de sus experimentos y el análisis de sus resultados.

2. El diseño de un experimento

> *Dame acierto al empezar, dirección al continuar, perfección al acabar.*
> Sto. Tomás de Aquino

Es imposible llevar a cabo un buen experimento en docencia que use de la estadística sin un cuidadoso diseño. El libro fundacional y más clásico de diseño de experimentos es The Design of Experiments, de R.A. Fisher [6]. Identificaremos 2 errores comunes: la mala elección de la muestra y el exceso de inferencias.

2.1. El muestreo

En muchas universidades españolas, la encuesta de evaluación del curso se hace en línea: los alumnos que quieren se conectan a su cuenta del campus virtual y contestan. Muchas veces la contesta una minoría. A esto se llama una muestra autoseleccionada y todos sabemos que es poco representativa[1].

En lo que quizá no nos demos cuenta es que si al final del curso en el que hemos hecho un experimento pasamos una encuesta, o tomamos alguna medida, a *todos* los alumnos de la clase, esto sigue siendo una pequeña muestra y es fácil que esté mal obtenida.

La población de interés en el caso de nuestro experimento no son los alumnos de la clase, sino todos los alumnos de esta asignatura en el presente y en el futuro (ya que, si da buenos resultados, querremos seguir usando la innovación bajo estudio) o incluso toda la población universitaria española (ya que esperamos que algún otro profesor lea nuestra ponencia y aplique nuestros métodos). Por lo tanto nuestros 80 alumnos son en su conjunto una pequeña muestra.

Y es muy difícil obtener una buena muestra. Algunas dificultades pueden ser tenidas en cuenta por el profesor, otras, no.

[1] A menos que los resultados sean favorables.

Un caso típico es una asignatura de primero en el que hay varios grupos. Tenemos que tener en cuenta cómo se han asignado los alumnos a cada grupo. A veces la administración (o el programa de matrícula) asigna de forma aleatoria a los alumnos a los grupos. Este es el caso ideal y podemos suponer que los grupos son equivalentes. Otras, son los alumnos que escogen el grupo. Entonces típicamente los alumnos con buena nota de selectividad se matriculan antes y tienen tendencia a escoger el mismo grupo. Los grupos así creados son sesgados y al realizar cualquier diseño experimental, se tiene que tener en cuenta dicho sesgo.

Otro caso, éste más insidioso, sucede al tomar medidas o hacer encuestas al final del curso. La mala obtención de la muestra es debido a que no estamos contabilizando a los alumnos que han abandonado durante el curso. Este efecto se conoce como *falacia de Neyman*. Supongamos que nuestro nuevo método da lugar a una calificación media inferior a la tradicional. Y además la encuesta que hemos pasado muestra que hay un menor grado de satisfacción entre los alumnos. Si miramos sólo esto podríamos pensar que el nuevo método es malo y debe abandonarse. Pero también tiene una menor tasa de abandono. Ahora otra interpretación de los resultados es posible: el método anterior sólo iba bien a cierto tipo de alumnos. A esos alumnos les gustaban mucho el método y obtenían buenas notas. Pero los demás abandonaban. Con el nuevo método alumnos que antes debían abandonar, ahora acaban y aprueban. Quizá no obtienen notas maravillosas y no estén tan contentos. Pero acaban, y eso es mejor que el abandono. ¿Cuál de las dos interpretaciones es la correcta? Eso requiere un análisis más completo. Hay que incluir las tasas de abandonos, comparar la distribución de las notas en ambos métodos y algunas cosas más. O incluso podemos comparar la evolución académica en cursos posteriores de aquéllos que han pasado por la innovación docente y aquellos que no. Un análisis simple, como comparar la media final, no basta.

2.2. A ver qué encuentro

A menudo cuando diseñamos un experimento, no tenemos claro qué es lo queremos averiguar ni las medidas que debemos tomar para averiguarlo. Pero el tiempo apremia y sólo podemos hacer un experimento por curso. El sentido común nos dice que debemos coger información de todas las variables que pensemos que puedan resultar relevantes y, una vez tengamos los datos, ya veremos cómo los vamos a usar. Y siguiendo esta misma lógica, cuantas más inferencias realicemos, cuanto más analicemos los datos, mejor.

Este sentido común tiene su razón de ser fuera del mundo de la estadística. Si no hay incertidumbre y las verdades están allí para ser descubiertas, cuántos más variables consideremos y cuánto más busquemos, me-

jor. En el uso de la estadística, dicha actuación es un grave error.

Para exponer mejor el error de este procedimiento supongamos que la "innovación" es tirar una moneda al aire 10 veces y tenemos una mejora si salen más caras que cruces. Entendemos que sólo sacar 6 caras puede no significar nada, por lo tanto consideraremos que la mejora es "de verdad" si salen 8 o más caras.

Damos una moneda a cada alumno y les pedimos que la lancen al aire 10 veces en cada sesión del curso y que apunten el resultado. Como no sabemos qué puede influir, recogemos datos de muchas variables: sexo, edad, nota de selectividad, si es diestro o zurdo, si lanzó la moneda en la primera mitad de clase o la segunda, el color de calcetines, el signo del zodiaco, etc. Una vez recogidos los datos nos ponemos a analizarlos. Lo extraordinario sería que no apareciera ninguna correlación. Seguramente encontraremos algo del estilo «Esta innovación produce una mejora estadísticamente significativa en los aries que llevan calcetines azules». Un caso real de búsqueda de correlaciones absurdas puede verse en un artículo de C. Aschwanden [3].

Es obvio que la causa de esta correlación no es el signo de zodiaco ni el color de los calcetines, sino la variabilidad natural del experimento y que la correlación es pura casualidad y falsa. Dicho error es muy común y es la base de la crisis de la replicabilidad [9, 13] que asola a muchos campos (la educación entre ellos).

La incertidumbre inherente a la estadística hace que haya los molestos pero inocuos falsos negativos, afirmaciones que son ciertas pero que no se pueden deducir de los datos obtenidos, y los peligrosos falsos positivos, afirmaciones falsas pero que sí se deducen de los datos. La probabilidad de un falso positivo es baja, pero si buscamos y buscamos y buscamos la probabilidad de que nuestra afirmación sea un falso positivo puede ser muy alta.

Un experimento como el indicado puede servir para hacer una exploración previa —que es inválida y de la que no usaremos los datos—, para entender mejor lo que pasa y así poder diseñar correctamente el experimento válido. En nuestro ejemplo, tras la exploración, dado que queremos estudiar si la innovación de la moneda ocurre entre los estudiantes, y en particular entre los aries y los que llevan calcetines azules, debemos dividir nuestra clase en cuatro grupos de tamaños similares, los aries con calcetines azules, los aries sin calcetines azules, los que llevan calcetines azules y no son aries y los que ni son aries ni llevan calcetines azules. A cada alumno del grupo le hacemos lanzar la moneda al aire un número determinado de veces y anotamos los resultados. Y ahora, usando las adecuadas estadísticas que hemos determinado con antelación, realizamos el análisis. Y, a pesar de lo que nos ha salido en el experimento exploratorio, va a ser que no, que ni los calcetines ni el signo zodiacal influyen.

3. Estadística descriptiva

La estadística descriptiva se encarga de ayudarnos a crear una historia del conjunto a partir de todos los datos individuales. Quizá los dos personajes más conocidos de esta historia son la media y el porcentaje. Dos simples números que usamos a menudo como representación de todo el conjunto. Una representación a menudo *demasiado* simple.

3.1. Una imagen vale más que mil medias

La mayoría de las técnicas estadísticas usuales, por no decir todas, se basan en el cálculo de estadísticos a partir de los datos de los que se dispone, sin prestar demasiada atención a los gráficos de los mismos.

El cuarteto de Anscombe [1] es un conjunto de 11 valores correspondientes a cuatro pares de variables $X_n^{(k)}, Y_n^{(k)};\ n = 1, \ldots, 11,\ k = 1, 2, 3, 4$. Si calculamos las medidas habituales de los pares de datos, la media, desviación típica, la correlación entre cada x, y, la recta de regresión y el coeficiente de correlación de los cuatro pares, sus valores son idénticos. Son indistinguibles. Pero si miramos las gráficas de la figura 1, vemos que son claramente diferentes.

No, los números no bastan. Hemos de mirar todos los números que podamos y todas las gráficas que se nos ocurran para hacernos una idea de qué representan nuestros datos.

3.2. La media y la mediana

«La nota media del curso fue de 7,2». «Los alumnos entregaron de media 9,3 problemas». «Cada alumno escribió de media 21,2 entradas al foro». Estas son frases típicas de muchos trabajos de docencia. La media es, con mucho, el número que usamos más a menudo como representante de un conjunto de datos numéricos.

Y aunque *sabemos* que esto no significa que cada uno de los alumnos obtuvo una nota media de 7,2, o que cada alumno escribió 21 entradas en el foro, sí que creemos en nuestro interior que cada alumno obtuvo una nota de *más o menos* 7,2 o que cada alumno escribió *más o menos* 21 entradas la foro. Volvemos a intentar concentrar en un dato simple la complejidad e incertidumbre del conjunto.

A veces es así. Si los datos son unimodales, simétricos y no tienen valores atípicos, es decir, si son muy poco problemáticos, la media y la desviación típica nos pueden dar una representación adecuada del conjunto. Pero no es raro tener datos problemáticos.

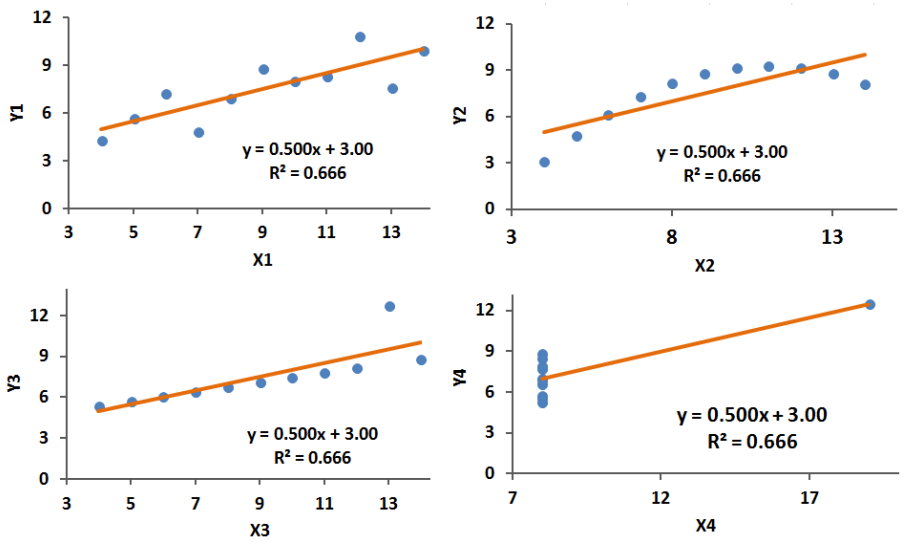

Figura 1: El cuarteto de Anscombe.

Reestudiemos el caso del foro, con las 21,2 entradas de media por alumno. Puede pasar que tengamos media docena de alumnos que no han entrado nunca, una mayoría de alumnos que hayan escrito unas pocas entradas, digamos que entre 5 y 10, y 2 alumnos enganchados al foro que tienen centenares de mensajes cada uno. Vemos que, en este caso, los 21,2 mensajes de media no representan, ni siquiera aproximadamente, a nadie: hay alumnos que escriben unos 10 mensajes, otros escriben cientos, pero nadie escribe cerca de 21 mensajes. Si sólo tenemos la media, cuando en el curso siguiente tengamos un media de 6,3 mensajes por alumno nos preguntaremos qué ha pasado. La respuesta muy probablemente es que no ha pasado nada. Simplemente este año no tenemos a los "forófilos".

Cuando nuestros datos son problemáticos —quizá fuertemente asimétricos, quizá con algún valor atípico— la media es un mal representante del conjunto. Una posibilidad es usar la mediana, mucho menos sensible a los valores extremos. En el ejemplo anterior la mediana estará alrededor del 7, lo que es mucho más representativo. Otra posibilidad es describir los datos mediante la distribución completa, el histograma y el *boxplot*, junto con cuartiles y quizá algún otro dato numérico. Es menos simple, tanto de describir como de calcular, como de entender, pero es una mejor descripción de los datos y de la incertidumbre que incluyen. Por tanto, antes de realizar un análisis numérico de los datos, es aconsejable realizar un gráfico de los mismos para tener una idea de cómo se distribuyen, ya sea un diagrama de caja o un histograma, o los dos. Y después decidir si la media es una buena medida global, o si necesitamos un conjunto de estadísticos para describir nuestros datos.

4. Estadística inferencial

La estadística inferencial es la que nos ayuda a realizar razonamientos estadísticos. Las pruebas y técnicas con las que nos provee nos da información, y en particular probabilidades, que nos ayudan a establecer la validez de los razonamientos que hagamos. Es importante destacar otra vez que la estadística vive en la incertidumbre y que los razonamiento son por lo tanto probabilísticos. La estadística inferencial no nos dice si algo es cierto o falso, no nos da oráculos. Olvidar esta afirmación crea una falsa seguridad que da lugar a comportamientos erróneos.

4.1. Intervalos de confianza

Lees un artículo sobre un nuevo método docente que se ha aplicado en muchas universidades en varias asignaturas. En dicho artículo, el autor afirma que en conjunto se ha conseguido una tasa de aprobados del 67 %, con un intervalo de confianza de [65 %, 69 %] a un nivel de confianza del 95 %. Tú impartes una asignatura similar a las del artículo. Tu tasa de aprobados es estable y se mueve alrededor de 58 %. Decides cambiar de método porque claramente es mejor que el tuyo, al menos en lo que respecta a la tasa de aprobados, pues tienes un 95 % de probabilidad de que esté entre el 65 % y 69 %. Y aunque no fuera ni el 65 %, seguro que es mejor que el 58 % que obtienes ahora. Desgraciadamente para ti, la interpretación que has realizado del intervalo de confianza es errónea.

Para ilustrar dicho error, pongamos un sencillo contraejemplo. Supongamos que tienes 10 alumnos en clase. Entonces la probabilidad de que tu tasa de aproba-

dos esté entre el 65 % y 69 % es 0.

Lo que nos dice el intervalo de confianza es que la tasa de aprobados de todos los alumnos que utilicen ese método en el pasado, presente y futuro está con un 95 % de probabilidades entre el 65 % y el 69 %. Y con una probabilidad del 5 % está fuera. La tasa que obtengas, suponiendo que uses exactamente el mismo método y las condiciones de tu clase sean las mismas que las del experimento, contribuirá a que ese total mundial y eterno se acerque a un número que es muy probable que esté dentro del intervalo [65 %, 69 %]. Pero tu tasa puede que siga siendo el 58 % de siempre.

Veamos una interpretación que se acerca más a la realidad, aunque tampoco es estrictamente cierta. Utilizas el nuevo método en tu asignatura y hallas la proporción de aprobados sobre matriculados. Calculas el intervalo de confianza de tú experimento. Hay una alta probabilidad que el 67 % del artículo esté dentro de *tu* intervalo.

4.2. El p-valor como oráculo

Queremos comercializar un suplemento alimenticio para reducir el colesterol llamado Sanacol. El departamento de publicidad quiere poder decir que tomándolo durante un mes el colesterol se reduce de media un 10 %. Si no llega a este 10 %, el éxito comercial es dudoso. Escogemos una muestra de personas con colesterol alto y sus niveles de colesterol son de 221, 235, 208, 214, 235, 201, 222, 248 y 218. Después de un mes de tomar Sanacol sus niveles se han reducido a 180, 205, 184, 193, 191, 185, 193, 212 y 204. Hacemos un t-test con el porcentaje de reducción de cada individuo y nos sale un p-valor de 0,053, por encima del nivel de significación de 0,05. ¡Rayos!, no es estadísticamente significativo y la campaña de publicidad se nos ha ido al garete. Pero, remirando los análisis con el técnico de laboratorio, vemos que el nivel de colesterol del último era un poquito menor que 204. Quizá incluso 203. Rehacemos el t-test y, ¡albricias!, el p-valor ha pasado a valer 0,047, por debajo del nivel de significación. Podemos dar el visto bueno a la campaña de publicidad.

Una persona que no sepa nada de estadística, pero con un módico de sentido común inmediatamente se da cuenta que aquí hay algo que no funciona. ¿Qué es este asombroso p-valor que establece verdades de una forma tan mágica?

Esta situación absurda proviene de uno de los errores más habituales en estadística: el considerar que el p-valor establece la verdad o falsedad de una hipótesis, y que además, esta verdad está relacionada con un umbral, el nivel de significación.

El problema no está en el p-valor ni en su definición. El p-valor es una probabilidad, la probabilidad, suponiendo que la hipótesis nula es cierta, de obtener valores como los de la muestra o más extremos. Es lógico que si cambia un poquito un valor, cambie un poquito la probabilidad. El problema es que no queremos tratar el p-valor como una probabilidad, sino que lo queremos usar como un oráculo, que nos diga si la hipótesis nula es cierta o falsa. Introducimos para ello el nivel de significación, que en el fondo es un valor arbitrario, que nos crea un falso umbral que nos permite pasar del cierto al falso. Esto da lugar al comportamiento tan estúpido que acabamos de describir. Y esto es aún peor si el valor de significación se decide *después* de calcular el p-valor: si nos sale $0{,}008$ decimos que es significativo al nivel de significación $\alpha = 0{,}01$ pero si nos sale mayor, digamos $0{,}032$, pues sigue siendo significativo, pero al nivel de significación $\alpha = 0{,}05$. La deshonestidad de este tipo de razonamiento es aparente. Precisamente estas búsquedas despiadadas de p-valores menores que $0{,}05$ (o $0{,}01$ o el valor que sea) contribuye enormemente a la crisis de la replicabilidad [13] que asola a muchos campos (la educación entre ellos). Por este motivo, la American Statistical Association sacó en marzo de 2016 una declaración con consejos que limitan el uso de los p-valores.

El p-valor es la probabilidad de obtener valores como los que hemos obtenido o más extremos suponiendo que la hipótesis nula es cierta. Y aunque la definición es corta y parece simple, incluso los expertos tienen dificultades para explicarla a un lego [2]. Los errores de interpretación cuando hay un p-valor, abundan.

Podemos establecer un símil de un estudio estadístico con la realización de un juicio. Por ejemplo, si el estudio estadístico consiste en averiguar si un método de enseñanza A es mejor que otro método de enseñanza B, en un juicio, vendría a ser ver si un acusado es culpable o inocente. Ser culpable equivaldría a afirmar que el método A es mejor que el B y ser inocente, que los dos métodos son igualmente eficaces. Declarar al acusado inocente o culpable depende de las pruebas de las que disponga el fiscal. Volviendo al estudio estadístico, la inferencia estadística sería la "oficina del fiscal" encargada de proporcionar dichas "pruebas". Obtener un intervalo de confianza o obtener un p-valor son simplemente pruebas de "acusación" o "inocencia" pero no bastan por sí solas para tomar la decisión final de inocencia o culpabilidad. En el juicio se pueden cometer dos errores: declarar culpable a un inocente o declarar inocente a un culpable. El primer error está controlado por el nivel de significación y sería desde el punto de vista ético el más grave. Aquí es donde el p-valor interviene de forma más directa ya que representa el nivel de significación máximo que estamos dispuestos a asumir para declarar culpable a un inocente. El segundo error está controlado por la potencia del contraste estadístico. Cuánto más poten-

cia tenga el contraste, más difícil sería cometer dicho error. En resumen, lo deseable para un contraste o estudio estadístico sería tener un nivel significación bajo y una potencia alta. Desgraciadamente, cuánto menor es el nivel de significación, menor es la potencia y al revés, si aumentamos la potencia del contraste, el error de significación aumenta.

Un p-valor muy pequeño quiere decir que es poco probable obtener datos como los que hemos observado. Desgraciadamente muchos creen que esto quiere decir que podemos rechazar con bastante seguridad que la hipótesis nula sea cierta, o, en otras palabras, que si rechazamos la hipótesis nula, es poco probable que nos hayamos equivocado. En el blog de F. Schönbrodt [11] muestran como, con unas suposiciones bastante razonables, la probabilidad de equivocarse al rechazar una hipótesis basado en un p-valor con un nivel de significación de $\alpha = 0{,}05$ puede ser del 25 %.

Hay más interpretaciones comunes y falsas que surgen de los p-valores. Una es que si el p-valor es muy pequeño, significa que la hipótesis nula es "muy" falsa o que los valores reales están muy alejados de los que hemos supuesto en la hipótesis nula. Supongamos que con un método la media de las calificaciones de aprobados es de 6,81 y con otro, de 6,84. Esta diferencia tan pequeña entre las medias no nos dice mucho sobre si un método es mejor que el otro. Incluso si el p-valor fuera $0{,}00002$, no nos diría gran cosa. Seguramente hay muchas otras cosas a comparar que nos dará mucha más información sobre qué método es mejor que esa pequeña diferencia de medias.

Podemos obtener p-valores minúsculos aunque la diferencia entre los valores hipotetizados y los reales sean insignificantes: todo depende de la varianza y del tamaño de la muestra. Y hoy en día, con el "big data" es cada vez más fácil obtener diferencias que son a la vez insignificantes y estadísticamente significativas.

Un p-valor es sólo un dato más que añade o reduce la seguridad que tenemos en nuestros razonamientos. Un dato más que hay que añadir a lo que hemos observado de los datos, los intervalos de confianza, el tamaño del efecto y otras pruebas que hayamos hecho, los razonamientos no estadísticos, el sentido común. Pero por alguna extraña razón un p-valor minúsculo ejerce una poderosa fascinación que nos hace pasar por alto los demás datos. Son errores tan antiguos como los p-valores y que el mismo Fisher, creador del concepto, trató en vano de corregir [7].

5. La visión bayesiana

La visión y métodos que hemos descrito de inferencia estadística corresponden a la visión clásica (o frecuentista). Fueron desarrollados por los pioneros de la estadística a principios del siglo pasado. Eran una primera solución, con los problemas de que hemos descrito y otros muchos. Los mismos creadores de los métodos —Fischer, E. Pearson, Neyman— eran conscientes de los problemas pero esperaban que en el futuro aparecieran nuevos métodos mejores.

Y efectivamente apareció otra visión y otros métodos: el enfoque bayesiano. Pero es un enfoque más complejo, con dificultades de cálculo y una interpretación de resultados más directa pero menos simple. Quizá sea por eso que no ha sustituido al enfoque clásico, que es el que siguen usando generalmente en los artículos en medicina, psicología y docencia. También es la única visión que se explica en la mayoría de los cursos de estadística. Y es por eso que nosotros nos hemos centrado en ella.

Pero creemos que es importante que el investigador en docencia conozca la existencia del enfoque bayesiano y de su fundamento. Por eso dedicamos esta sección a ilustrar sus conceptos y funcionamiento. Para ello vamos a resolver un problema simple de forma bayesiana.

En la visión clásica el objetivo es responder a la pregunta «¿Qué puedo saber con estos datos?». El objetivo, y la pregunta, en el enfoque bayesiano es diferente: «¿Cómo cambian estos datos mi conocimiento de la situación?». Esto implica que en el enfoque bayesiano tengo un conocimiento previo, la distribución *a priori* y quiero obtener un nuevo conocimiento, la distribución *a posteriori*. El método se puede resumir, de manera esquemática, en los siguientes pasos: i) definir la distribución *a priori* con la información que se tenga disponible (previa informativa) o reflejar en ella desconocimiento del proceso (previa no informativa); ii) describir la verosimilitud de los datos observados; iii) usar el teorema de Bayes para calcular la distribución *a posteriori*; y iv) con la distribución completa *a posteriori* calcular lo que nos interese: la estimación de su valor, los intervalos de credibilidad, e incluso la probabilidad de que ciertas hipótesis sean ciertas o falsas, cosas que no se puede hacer con la visión frecuentista.

Veamos cómo se calcula un intervalo de confianza con el enfoque bayesiano. Aquí sólo mostraremos un esquema en el caso de una distribución discreta. En `http://bioinfo.uib.es/~joemiro/bayesiano` está la resolución detallada con todos los cálculos e incluyendo el caso de una distribución continua.

Supongamos que hemos hecho una prueba con 100 alumnos y 67 la han superado. Esto nos da un estadístico para la tasa de aprobados $\hat{\theta} = 0{,}67$. El método clásico parte del teorema central del límite que nos dice que la distribución de $\hat{\theta}$ es normal, y de aquí obtenemos fácilmente que el intervalo de confianza con un nivel de confianza del 95 % es $[0{,}58, 0{,}76]$. No sabemos cuánto vale la tasa "real" θ pero en el 95 % de los casos el

método nos dará un intervalo que lo incluya.

Para calcular el intervalo de forma bayesiana debemos partir de una distribución *a priori* de θ. Como no sabemos nada todavía, vamos a suponer que la tasa es del 50 %. Esto no nos basta: necesitamos una distribución completa para θ que refleje nuestra incertidumbre. Elegimos la siguiente:

θ	0.3	0.4	0.45	0.5	0.55	0.6	0.7
Prob	0.05	0.1	0.15	0.4	0.15	0.1	0.05

A partir de estas probabilidades *a priori*, usamos el teorema de Bayes para calcular las probabilidades *a posteriori*. El teorema nos dice que con 67 aprobados, para cada valor de θ de la distribución *a priori*, la probabilidad *a posteriori* es

$$Pr(\theta \,|\, 67\, ap) = \frac{Pr(67\, ap \,|\, \theta) \cdot Pr(\theta)}{\sum_{i=1}^{7} Pr(67\, ap \,|\, \theta_i) \cdot Pr(\theta_i)}$$

La probabilidad de $Pr(67\, ap \,|\, \theta)$ se calcula con la distribución binomial. Si calculamos la probabilidad *a posteriori* para $\theta = 0{,}45$ obtenemos una probabilidad de $3{,}06 \cdot 10^{-5} \approx 0$. Repetimos estos cálculos para todos los valores de θ y obtenemos la distribución *a posteriori* completa:

θ	0.3	0.4	0.45	0.5	0.55	0.6	0.7
Prob	0	0	0	0	0.03	0.26	0.70

A partir de esta distribución podemos calcular la estimación *a posteriori* de θ, que no es más que la esperanza de la distribución y obtenemos que es $0{,}668$. No hay una manera única de obtener el intervalo de credibilidad (el equivalente al intervalo de confianza). Escogemos el más estrecho que incluya a la moda. En este caso obtenemos un intervalo con un nivel de confianza del 96 %: $[0{,}6, 0{,}7]$

Si en vez de una distribución *a priori* discreta se desea partir de una continua, también se puede hacer, aunque los cálculos se vuelven más complejos.

El resultado de un análisis bayesiano es una distribución completa de probabilidad. Esto dificulta el responder en plan oráculo a nuestras preguntas, lo que es una buena cosa. Resuelve así algunos de los problemas que hemos descrito, aunque no todos [5]. Pero hay motivos por los que esta visión no ha sustituido a la frecuentista.

Un primer motivo es que la visión frecuentista es más fácil de explicar: todos entendemos la base frecuentista de definición de probabilidad: «casos favorables partido casos posibles». Explicar el concepto de probabilidad o las ideas básicas de la inferencia estadística desde el enfoque bayesiano requiere mayor conocimiento probabilístico y matemático. Este motivo, desgraciadamente, es autopropagante: muchos de los que usan estadística sólo conocen el método frecuentista, lo que lo hace el más usado, lo que obliga a enseñarlo primero (aunque sólo sea para poder entender los razonamientos habituales), con lo que sólo los que se adentran a cursos avanzados son expuestos a la visión bayesiana.

Un segundo motivo es que las conclusiones obtenidas suelen depender de la selección específica de la previa. ¿Por qué hemos usado esta distribución *a priori* y no otra? Hay una aparente carga de subjetividad en esta elección, y este enfoque, que la hace sospechosa ante el no experto.

Un tercer motivo es que la respuesta es una distribución completa y no un estadístico simple. Lo que es mirar un simple número en la visión frecuentista se convierte en estudiar una distribución. Esto, que es precisamente lo que reduce muchos de los problemas que hemos discutido, hace más compleja la interpretación de los datos.

Pero estos dos últimos motivos son, precisamente, lo que hacen interesante este enfoque. Es necesario hacer un estudio de robustez frente a la distribución *a priori* y esto permite al investigador tener diferentes escenarios para tomar una decisión. Aunque sea una visión interesante, supone también mucho trabajo, trabajo que, desde el punto de vista matemático, es difícil, sobre todo si se usan distribuciones *a priori* distintas de las existentes en la literatura. Para un no experto que usa la estadística como una más en su arsenal de herramientas, el enfoque frecuentista es mucho más atractivo que el bayesiano. Sobre todo si no es consciente de los problemas inherentes a él.

Un último motivo, la complejidad computacional, cada vez lo es menos, con la aparición de programas estadísticos como R [10], con paquetes como Learn Bayes o R2WinBUGS, o programas específicos como JAGS, *Just Another Gibbs Sampler*[2].

6. Conclusiones

A todos nos gusta la certeza. Y a todos nos gusta lo fácil y simple. De aquí la enorme atracción de la idea siguiente: «tomo unas medidas, las meto en un ordenador, aprieto una tecla, miró un número y establezco con casi total seguridad si un método docente es mejor o peor que otro». En el fondo sabemos que la vida no es así y que esto no puede ser cierto. Pero el ansia de certeza puede engañarnos una y otra vez, incluso a los expertos en estadística.

La estadística nos permite gestionar la incertidumbre. Nos permite diseñar los experimentos de manera que reduzcamos la probabilidad de tener falsos positivos y que podamos usar argumentos rigurosos. La estadística descriptiva nos permite hacernos una idea intuitiva de qué representan los datos y de lo problemáticos que son, ayudándonos a decidir qué pruebas debemos

[2]Disponible en `http://mcmc-jags.sourceforge.net/`

aplicar y qué se puede deducir de ellas. La estadística inferencial nos calcula probabilidades que nos permite estimar la incertidumbre de nuestros resultados. Todo junto permite al investigador cuidadoso afirmar con cautela sobre la probable veracidad de las hipótesis. La estadística sólo permite afirmaciones del estilo «Si las medias fueran iguales, es poco probable obtener datos como estos». Para pasar de aquí a «el método A es mejor que el B» deben añadirse razonamientos basados, por ejemplo, en teorías del conocimiento. La estadística sola no puede llegar allí.

El conocimiento de la estadística es imprescindible para cualquier investigador en educación. La naturaleza misma del campo, con su incertidumbre y variabilidad, lo requiere. Pero no es un conocimiento de cómo realizar cálculos o cuáles son las pruebas habituales. Ese es el conocimiento fácil y que casi todo el mundo tiene. Lo que necesitamos es saber cómo realizar un razonamiento estadístico y cómo interpretar los resultados. No es una tarea sencilla.

Agradecimientos

Este trabajo está parcialmente subvencionado por el projecto de investigación del Ministerio de Economía y Competitividad con referencia DPI2015-67082-P.

Referencias

[1] F. J. Anscombe. *Graphs in Statistical Analysis.* The American Statistician, vol. 27, núm. 1, pp. 17–21. Febrero de 1973.

[2] Christie Aschwanden. *Not Even Scientists Can Easily Explain P-values*. Five ThirtyEight.com. http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/. Noviembre de 2015. Fecha de último acceso: febrero de 2017.

[3] Christie Aschwanden. *You Can't Trust What You Read About Nutrition*. Five ThirtyEight.com. http://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/. Enero de 2016. Fecha de último acceso: febrero de 2017.

[4] William M. Briggs. *Breaking the law of averages. Real-life probability and statistics in plain English.* 2008. Disponible en http://wmbriggs.com/public/briggs_breaking_law_averages.pdf

[5] William M. Briggs. *Uncertainty: The Soul of Modeling, Probability & Statistics.* Springer. 2016.

[6] Sir Ronald A. Fisher. The design of experiments, 9ª edición. Harper Press, Londres 1971.

[7] Jim Frost. *Why Are P Value Misunderstandings So Common?* The Minitab blog. 10 de diciembre de 2015. Disponible en http://blog.minitab.com/blog/adventures-in-statistics-2/why-are-p-value-misunderstandings-so-common. Fecha de último acceso: febrero de 2017.

[8] Daniel Kahneman, Paul Slovic y Amos Tversky (eds.). *Judgment under uncertainty: Heuristics and biases.* Cambridge university press. 1982.

[9] Open Science Collaboration. *Estimating the reproducibility of psychological science* Science, vol. 349, Issue 6251, 28 Aug 2015.

[10] R Core Team, *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. 2016.

[11] Felix Schönbrodt. *What's the probability that a significant p-value indicates a true effect?* Blog Nicebread, 3 de noviembre de 2015. Disponible en http://www.nicebread.de/whats-the-probability-that-a-significant-p-value-indicates-a-true-effect/ Fecha de último acceso: febrero de 2017.

[12] Ronald L. Wasserstein y Nicole A. Lazar. *The ASA's Statement on p-Values: Context, Process, and Purpose*. The American Statistician, vol. 170, núm. 2, pp. 129–133. 2016.

[13] Kirsten Weir. *A reproducibility crisis?* Monitor on Psychology, vol. 46, núm. 9. Octubre de 2015.