EL PAPEL DE LA ESTADÍSTICA EN LA INVESTIGACIÓN CIENTÍFICA

A. SANTANA

1. Un poco de historia

Podemos retrotraer los orígenes de la estadística a las actividades de organización administrativa de las antiguas civilizaciones. Egipcios, babilonios o sumerios en épocas tan remotas como el año 3000 A.C. debieron enfrentarse al problema de recopilar información que les permitiera saber con qué recursos contaban para hacer funcionar sus sociedades: ¿cuántos hombres podían formar parte del ejército? ¿Cuánto terreno agrícola había disponible? ¿Sería suficiente la cosecha para alimentar al pueblo? ¿Cuando se produciría la próxima crecida del Nilo y hasta donde alcanzaría? ¿Cuántas cabezas de ganado había? ¿Cuántos hijos nacían cada año? Y la pregunta clave ¿Cuántos impuestos se podían cobrar para poder financiar el estado? Téngase en cuenta que en esta época, si bien los servicios públicos eran escasos (en muchas ocasiones, nulos), había que hacer frente al pago de las campañas militares, la construcción de palacios, templos, conducciones de agua, etc.

En realidad todas las sociedades más o menos organizadas a lo largo de la historia han debido dotarse de medios para conocer su situación socioeconómica: en Europa la antigua Grecia, el imperio romano, los reinos medievales, las repúblicas renacentistas, ...; en Asia, el imperio chino o el imperio mongol; en América los incas, mayas y aztecas; el imperio musulmán en el norte de África, ... En cualquier caso, estas recopilaciones de información socioeconómica no se hacían bajo un criterio universal, siguiendo unos procedimientos más o menos normalizados y comparables entre los distintos estados, reinos, naciones o imperios, sino que cada uno de ellos se las ingeniaba de la mejor forma que podía.

Fue en 1690, cuando el economista inglés William Petty trató por primera vez en su libro *Aritmética Política* de establecer un cierto orden en el modo en que los estados (en su caso muy concreto, la corona inglesa) debían llevar el registro de los medios con que contaban (tierras, personas, edificios, ganadería, puertos, barcos, impuestos, ...). Casi 60 años más tarde, en 1749 el economista alemán Gottfried Achenwall utilizó por primera vez la palabra *"Statistik"* en su libro *"Resumen de la más reciente Ciencia Estatal de las principales naciones europeas y repúblicas para uso en sus Conferencias Académicas"*, refiriéndose con este término al estudio de la situación del estado. Así, a mediados del siglo XVIII la palabra "estadística" empezó a usarse para referirse a la ciencia que se ocupaba de reunir, sintetizar y analizar de una manera más o menos estandarizada los datos del estado. No obstante, no pasó mucho tiempo para que este término pasara a englobar cualquier proceso de recopilación y análisis de datos también en los ámbitos de la ciencia o la actividad empresarial. En esa época, ya hacía tiempo que la recopilación de datos había dejado de ser patrimonio exclusivo del estado. Así, por ejemplo, las nacientes compañías de seguros hacían uso extensivo de los datos demográficos que ellas mismas recopilaban; muchos matemáticos de entonces mostraron interés por el estudio de las *tablas de vida* (cuántas personas vivían hasta ciertas edades, cuál era la mortalidad infantil, cuál era el riesgo de morir por diversas causas,...); las compañías de las Indias (Occidentales y Orientales) llevaban extensos registros de sus actividades, con el claro interés de ajustar su capacidad comercial a la demanda de productos; los astrónomos llevaban siglos compilando extensas tablas de datos sobre la posición de estrellas, planetas o cometas. En resumen, cuando Achenwall puso nombre a la Estadística, en realidad estaba "bautizando" a una criatura que había nacido hacía ya tiempo y que se encontraba en pleno crecimiento.

Mientras tanto, y de manera absolutamente desconectada de ese concepto de estadística como ciencia de los datos del estado, una plétora de estudiosos europeos en distintos países (Italia, Suiza, Países Bajos, Inglaterra, Francia, Alemania) había iniciado el estudio de la probabilidad, muy relacionado en principio con el estudio de los juegos de azar y las apuestas. Una de las primeras referencias en este sentido la encontramos en una carta remitida por Pascal a Fermat en 1654 donde le expone un problema planteado por un amigo suyo, el *caballero de Meré*. El problema en concreto consistía en decidir si era mejor apostar a que salga al menos un 6 cuando se tira un dado cuatro veces, o apostar a que salga al menos un doble 6 cuando se tiran dos dados 24 veces. En realidad el planteamiento de este problema estaba probablemente inspirado en los datos recogidos por el propio Meré durante sus incursiones en las casas de apuestas de la época (figura 1).

Alentados por problemas como éste, y por su propia curiosidad, matemáticos como De Moivre, Fermat o Laplace construyeron entre los siglos XVII y XVIII el edificio teórico en que se sustenta la teoría de la probabilidad. Y aunque en principio la probabilidad podía parecer casi un divertimento dedicado al estudio de los juegos de azar, no pasó

FIGURA 1. En la Francia del siglo XVII los juegos de dados, cartas y tableros con fichas eran los entretenimientos más frecuentes. Los juegos, cada vez más complicados, y el interés por ganar con apuestas cada vez más elevadas estimularon el interés por el estudio de la probabilidad.

mucho tiempo antes de que los matemáticos que se dedicaban a su estudio empezaran a encontrar muchas otras aplicaciones prácticas. Una de tales aplicaciones, quizás de las primeras que conectó el mundo de la estadística (la recopilación de datos), con la probabilidad vino del ámbito de la Astronomía. En el siglo XVIII la mejor y mayor disponibilidad de instrumentos ópticos empezó a plantear una serie de problemas nuevos a los astrónomos de la época, como calcular la órbita de ciertos cometas, explicar ciertas irregularidades en las órbitas de Júpiter y Saturno, o incluso determinar de manera lo más precisa posible del movimiento de *libración* de la Luna (un cierto movimiento oscilatorio que hace que aunque la Luna muestre siempre la misma cara hacia la Tierra, desde nuestro planeta podamos ver algo más de la mitad de la superficie lunar). El gran problema es que las observaciones no eran precisas y los astrónomos eran conscientes de que entre la posición observada (de la Luna, de un cometa, de Júpiter o de Saturno) y la posición real había siempre una diferencia, el *error experimental* o *error de observación*. En el contexto de este problema, Legendre y Gauss desarrollaron de manera independiente el **método de los mínimos cuadrados**, que permitía aproximar la órbita de un cuerpo celeste a partir de una colección de observaciones con error (figura 2).

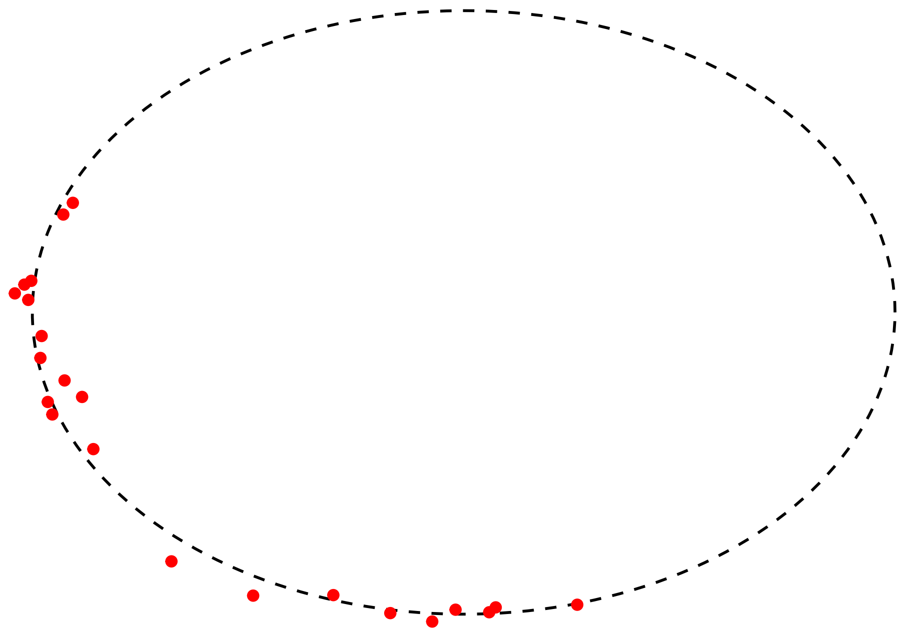

Figura 2. Dado un conjunto de observaciones de la posición de un planeta, el método de mínimos cuadrados permite determinar la ecuación de la elipse que mejor se ajusta a dicho conjunto de observaciones

Mientras que Legendre (1805) se limitó a resolver el problema de manera algebraica, Gauss (a partir de 1809) desarrolló además la teoría estadística de errores, de la que el método de mínimos cuadrados resultaba como un caso particular. El interés de Gauss en este problema se suscitó en 1801, cuando el astrónomo Giuseppe Piazzi descubrió el planeta enano Ceres, que en ese momento se pensaba que podría ser un nuevo planeta (en la actualidad se sabe que forma parte del cinturón de asteroides). Desafortunadamente, después de aproximadamente un mes, la órbita de Ceres lo llevó a situarse detrás del sol. Como no se había recopilado suficiente información durante ese mes de observación no era posible predecir dónde reaparecería Ceres. De hecho, después del día en que se suponía que Ceres reaparecería, nadie pudo localizarlo. Sin embargo, Carl Friedrich Gauss, en aquel momento con veinticuatro años de edad, fue capaz de determinar en qué posición estaba, si bien en aquel momento no dijo como había podido hacer dicha determinación. Fue años después cuando explicó que para ello había considerado que los errores en las mediciones de Ceres no se repartían completamente al azar, sino que eran más probables los errores pequeños que los errores grandes; además un error de cierta magnitud por exceso era tan probable como un error con la misma magnitud por defecto. Bajo estos supuestos fue capaz de desarrollar un método que permitía encontrar cuáles eran los valores *más probables* para los parámetros que caracterizaban la órbita del objeto celeste.

Este trabajo de Gauss, combinando el uso empírico de datos observados con formulaciones matemáticas derivadas de la teoría de la probabilidad, fue una de las primeras aplicaciones prácticas de lo que hoy se conoce como estadística matemática: la ciencia que combina el uso de datos que proporcionan información incompleta sobre fenómenos de interés con técnicas probabilísticas que permiten utilizarlos para dar respuesta a preguntas complejas sobre dichos fenómenos.

Para terminar nuestro recorrido por la historia de la Estadística, debemos citar aquí a los autores que impulsaron a finales del siglo XIX y principios de XX el desarrollo definitivo de esta disciplina: Francis Galton, Karl Pearson, Ronald Fisher, Jerzy Neyman y Egon Pearson. Estos autores fueron los que dieron forma a lo que hoy se conoce como *inferencia estadística*: cuál es la mejor forma de utilizar los datos que proporciona una muestra particular para poder responder a cuestiones de carácter general. A partir de entonces la actividad científica quedó ligada permanentemente al uso de la metodología estadística.

2. El método científico

En la actualidad, las técnicas de inferencia estadística constituyen un pilar fundamental en el que se apoya el método científico. El funcionamiento del método científico, desarrollado por Galileo en el siglo XVI se sintetiza en la figura 3, y se ajusta al siguiente esquema:

1. En principio se realiza la observación de algún fenómeno de interés.
2. Se plantea una hipótesis que puede explicar como se produce dicho fenómeno.
3. Se diseña un experimento que permita comprobar la validez de dicha hipótesis.
4. Una vez realizado el experimento puede ocurrir que:
    a. El resultado del experimento confirme la validez de la hipótesis de partida. En este caso se comunica dicho resultado que, eventualmente, queda recogido en una *teoría científica* más amplia.

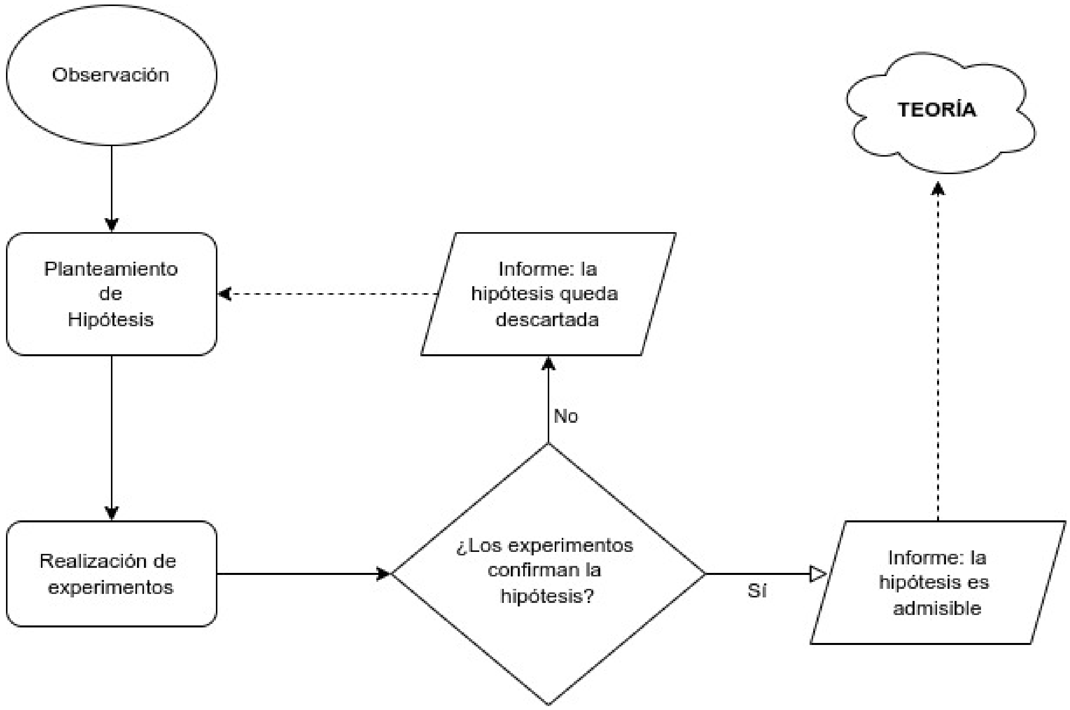

Figura 3. Esquema del funcionamiento del método científico.

b. El resultado del experimento no es compatible con la hipótesis planteada. En este caso, **la hipótesis queda descartada** y se puede proceder a repetir el ciclo, reformulando la hipótesis de otra manera y realizando nuevos experimentos.

Es de destacar que en el método científico las hipótesis *no quedan nunca demostradas*. Pueden confirmarse una y otra vez, pueden integrarse en un corpus de conocimiento sólido y consistente (lo que a grandes rasgos se denomina *teoría científica*). Pero siempre existe el riesgo de que algún experimento futuro dé un resultado que no pueda explicarse en el marco de la teoría, en cuyo caso ésta quedaría descartada como falsa. De hecho esta particularidad es la que caracteriza a las teorías científicas: no puede demostrarse que sean ciertas, pero existe la posibilidad de demostrar que son falsas. Una afirmación para la que sea imposible diseñar un experimento que eventualmente pueda demostrar que es falsa no es una afirmación científica. Como ejemplo podemos citar la teoría geocéntrica, que afirmaba que el Sol gira en torno a la Tierra. Esta teoría fue aceptada durante muchos siglos porque explicaba bien el ciclo día-noche; no obstante fue incapaz de explicar los movimientos de los planetas registrados por Kepler, por lo que dicha teoría decayó en favor de la teoría heliocéntrica. La teoría de la Gravitación Universal de Newton fue incapaz de explicar la posición de Mercurio durante un eclipse solar, cosa que sí explicaba la Teoría de la Relatividad de Einstein; por tanto, la primera fue descartada en función de la segunda (si bien los principios derivados de la gravitación de Newton se siguen usando porque para la mayor parte de aplicaciones prácticas de la física ofrecen una aproximación más que suficiente, al mismo tiempo que una formulación más sencilla que la teoría relativista).

¿De qué manera interviene la Estadística Matemática en el método científico que acabamos de describir? De varias formas:

1. En la observación del fenómeno. En fenómenos complejos es preciso realizar muchas observaciones, que se convierten en muchos datos. Las técnicas de la estadística exploratoria, utilizando tablas, gráficos y medidas de resumen permiten visualizar los resultados y apreciar las tendencias o regularidades que puedan estar presentes en los datos.

2. Una vez formulada la hipótesis de interés, la estadística ofrece métodos matemáticos para diseñar adecuadamente el experimento: cuántas muestras deben tomarse, de qué forma, ...

3. La estadística dispone de procedimientos para tomar decisiones teniendo en cuenta la variabilidad experimental. Un experimento es determinista si cuando se repite en las mismas condiciones produce siempre el mismo resultado. Sin embargo, esto no ocurre prácticamente nunca. En la práctica lo normal es que las distintas repeticiones del experimento produzcan resultados diferentes, lo que en muchas ocasiones complica decidir si los datos confirman o refutan la hipótesis puesta a prueba.

A continuación ilustraremos estas ideas con algunos ejemplos que nos ayudan a entender como se aplica la estadística al abordar cuestiones de interés científico.

Ejemplo 1: verificación de la ley de gravitación universal. En la Física clásica hay experimentos cuyo resultado tiene poco margen de variabilidad (son casi deterministas). Por ejemplo, cuando dejamos caer un objeto desde una altura determinada, el tiempo que tarda en llegar al suelo se puede medir con mucha precisión, y apenas difiere de lo que predice la fórmula de la gravitación. A continuación mostramos los datos resultantes de realizar diez veces el experimento de dejar caer una bola de plomo desde una altura $h = 10$ metros. De acuerdo con la ley de gravitación universal, el tiempo que tardará esta bola en llegar al suelo es $t = \sqrt{\frac{2h}{g}} = 1.429$ segundos. Los valores obtenidos en los 10 experimentos se muestran en la tabla 1.

TABLA 1. Tiempos observados en 10 experimentos en los que se ha dejado caer una bola de plomo desde una altura de 10 metros.

ensayo	**tiempo**
1	1.427
2	1.429
3	1.430
4	1.426
5	1.429
6	1.429
7	1.428
8	1.429
9	1.428
10	1.430

Como vemos, los resultados obtenidos apenas se apartan de la predicción hecha por la ley de la gravitación universal, por lo que no parece que haya razón para dudar de la veracidad de la misma (al menos para magnitudes de la escala observada en este experimento). Como en todos los ensayos no ha resultado el mismo valor, podemos asumir que el tiempo medido en el ensayo i es de la forma:

$$t_i = \tau + \varepsilon_i$$

donde τ es el tiempo de caída y ε_i es el error experimental en el ensayo i debido a múltiples factores (sincronización del cronómetro, rozamiento del aire, ...). El análisis estadístico de los datos nos indica que el valor de τ se encuentra *casi seguramente* entre 1.4276 y 1.4294. Por tanto, el valor que propone la ley de gravitación (1.429) se encuentra dentro del margen de error experimental y nada se opone a aceptar la validez de dicha ley.

Ejemplo 2: tratamiento para adelgazar. Imaginemos ahora que diez personas (de idéntica edad, peso, estatura, complexión física, sexo, etc) se presentan voluntarios para ensayar un nuevo tratamiento para reducir peso, consistente en la ingesta de una píldora antes de cada comida. Después de un mes sometidos al tratamiento, se mide la variación de peso observada (en kg) como la diferencia entre el peso final y el peso inicial. En la tabla 2 se muestra el resultado obtenido para cada uno de los sujetos experimentales. La figura 4 representa gráficamente estos resultados.

Tabla 2. Variación de peso (en kg.) experimentada a lo largo de un mes por diez personas que siguen un tratamiento para adelgazar.

sujeto	Variación Peso
1	0.127
2	-0.628
3	-1.521
4	-0.856
5	0.587
6	-1.976
7	0.194
8	-1.365
9	0.213
10	-1.116

Como vemos, cuatro de los voluntarios habían aumentado algo el peso inicial. Los otros seis habían experimentado una reducción de peso, con una disminución máxima

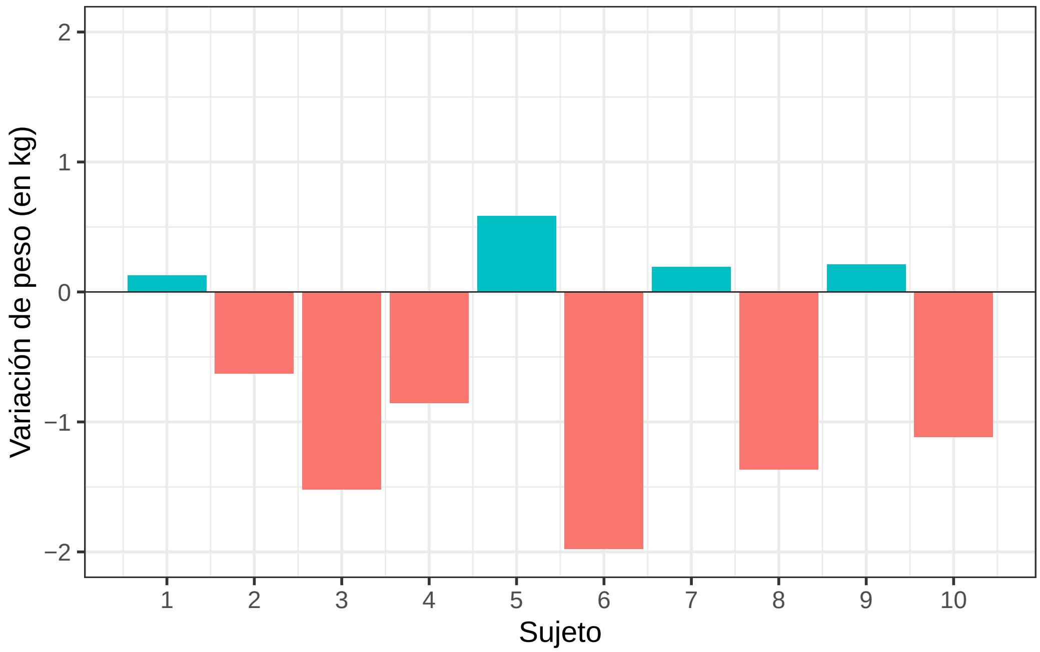

Figura 4. Representación gráfica de la variación de peso (en kg.) experimentada a lo largo de un mes por diez personas que siguen un tratamiento para adelgazar. En rojo se representa el peso perdido por las personas que han adelgazado y en azul el peso ganado por las que han engordado.

de casi 2 kg en el sujeto 6. Por tanto, aunque todos los sujetos partieron de las mismas condiciones iniciales y fueron sometidos al mismo tratamiento, hay una gran variabilidad entre las respuestas. Seguramente en la variación de peso experimentada por cada individuo han intervenido muchos más factores aparte del tratamiento: el metabolismo de cada uno, la actividad física diaria, diferencias en la dieta, … La suma de estos factores es lo que denominamos variabilidad aleatoria. Aún así, en este experimento parece que han sido más los sujetos que han conseguido adelgazar, y en mayor magnitud que el incremento de peso en aquellos que no lo han conseguido. Así pues, teniendo en cuenta todas estas circunstancias, con estos datos ¿puede decirse que el tratamiento es efectivo?

Un poco de reflexión sobre el problema nos lleva a darnos cuenta de que esta pregunta no tiene una respuesta clara. Todos sabemos que, sin ningún tratamiento especial, nuestro peso oscila de forma natural; así que la primera duda que nos asalta es: ¿estas variaciones de peso se pueden achacar realmente al tratamiento o se deberán simplemente a las oscilaciones naturales del peso de estas personas? Y si admitimos que podrían deberse al tratamiento: ¿el efecto de éste es realmente una disminución de peso? 6 sujetos disminuyeron de peso (una media de 1.244 kg), pero 4 aumentaron (una media de 0.28 kg). Además, solo hemos probado con diez voluntarios, lo que no parece que nos dé demasiada información.

Con estos resultados, nuestra intuición no sabe muy bien qué decisión tomar. Es muy posible que a algunas personas estos resultados le digan que el tratamiento puede ser efectivo. Y seguramente para otras personas estos resultados parecerán poco concluyentes. Ahora bien, si queremos tener una respuesta concreta a esta cuestión, no debemos fiarnos de nuestra intuición. Es aquí donde entra en juego la metodología estadística.

De entrada, la primera cuestión que hemos señalado es importante: no sabemos cuál es la oscilación natural del peso, así que no podemos saber si estas subidas y bajadas se pueden deber a la misma. Para poder tener una idea de la magnitud de dicha oscilación, lo mejor es tener un grupo control que no tome el tratamiento. Pero dicho grupo control debe a su vez *creer* que toma el tratamiento, para que sus resultados sean comparables a los de este grupo (sabemos que la mente es traicionera y la mera predisposición a conseguir un resultado puede terminar provocando que dicho resultado se consiga). Supongamos que tomamos a otros diez voluntarios de similares características que los anteriores, les decimos que les estamos dando el tratamiento, pero en realidad les damos un *placebo* (una pastilla similar a la del tratamiento, pero que no tiene efecto). La tabla 3 muestra los resultados obtenidos en este grupo de voluntarios, que se muestran gráficamente en la figura 5

Tabla 3. Variación de peso (en kg.) experimentada a lo largo de un mes por las diez personas que toman placebo (grupo control).

sujeto	Variación Peso
11	0.324
12	-2.261
13	-0.830
14	-1.629
15	1.938
16	0.078
17	-0.809

sujeto	Variación Peso
18	-1.867
19	1.524
20	0.683

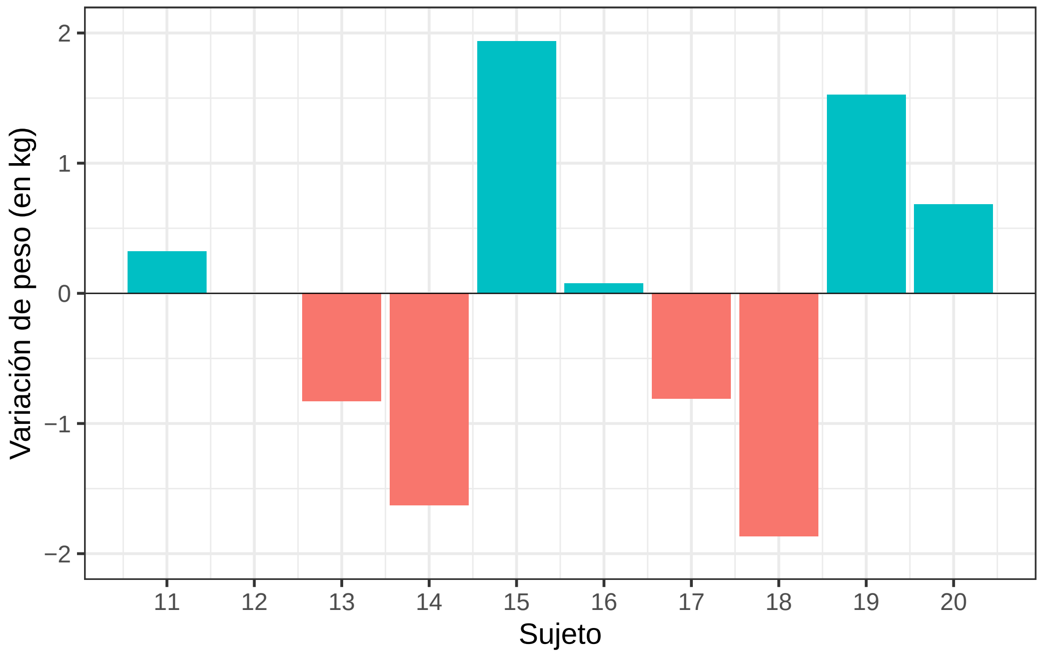

Figura 5. Representación gráfica de la variación de peso en el grupo control. En rojo se representa el peso perdido por las personas que han adelgazado y en azul el peso ganado por las que han engordado.

En este grupo los que han engordado, han engordado una media de 0.909 kg y los que han adelgazado lo han hecho una media de 1.479 kg. Si comparamos con el grupo anterior, en éste solo han adelgazado 4 personas, si bien en promedio adelgazaron más que los que tomaron el tratamiento. Asimismo, en este grupo han sido más los que han aumentado de peso, y en promedio han engordado más que los del otro grupo. ¿Podríamos concluir ahora que el tratamiento fue realmente efectivo? Seguimos sin tenerlo del todo claro. Vemos que los sujetos (tanto los del grupo tratamiento como los del grupo control) han experimentado subidas y bajadas de peso variables; en el fondo, todos son sujetos distintos, cada uno de su padre y de su madre, y aunque todos han seguido el mismo protocolo (tomarse una píldora antes de las comidas), todos han experimentado variaciones de peso distintas. Eso nos lleva a pensar: si hubiésemos reclutado 20 voluntarios distintos y hubiésemos repetido el experimento ¿habríamos obtenidos los mismos resultados? Con seguridad, no.

Hasta aquí hemos hecho estadística (hemos dibujado gráficas y calculado medias). Ahora es cuando sube al escenario la probabilidad. En este caso, el modelo adecuado para analizar estos datos es:

$$\Delta W_i = \mu + \tau \cdot T_i + \varepsilon_i$$

donde ΔW_i representa la variación de peso en el sujeto i, μ representa la variación media mensual de peso de los individuos de esta población, τ_i la variación de peso debida al tratamiento, T_i indica si el individuo ha tomado tratamiento (en cuyo caso vale 1) o placebo (en cuyo caso vale 0), y ε_i la variación de peso aleatoria debida a una multiplicidad de causas. Usando estos datos y este modelo podemos calcular la probabilidad del resultado obtenido, **bajo la hipótesis de que el tratamiento no tiene efecto**. Dicha probabilidad es aproximadamente 0.26. Téngase en cuenta que la probabilidad (no lo hemos dicho hasta ahora) es una medida de incertidumbre que toma valores entre 0 y 1; lo que tiene probabilidad 0 es imposible que ocurra y lo que tiene probabilidad 1 ocurrirá con seguridad. Una probabilidad 0.26 (o del 26%) ciertamente no es demasiado alta, pero tampoco es excesivamente baja. Los científicos se han puesto de acuerdo en considerar como muy bajas las probabilidades menores que 0.05. Cualquier probabilidad mayor que este valor se considera que corresponde a un evento que no es raro que suceda.

Por tanto, la conclusión de nuestro análisis es que el resultado obtenido en el grupo de sujetos que han tomado el tratamiento muy probablemente podía haberse observado también **si el tratamiento no tuviera ningún efecto**. Por tanto concluimos que el tratamiento **no ha demostrado** ser efectivo para bajar de peso. Y nótese que no decimos que el tratamiento no sea efectivo (a lo mejor lo es), sino que el experimento no ha podido demostrar que lo sea.

Obviamente, se podrá decir que no se ha conseguido demostrar la efectividad del tratamiento porque el experimento es poco informativo con solo 20 sujetos experimentales. Es posible que así sea, así que podemos preguntarnos ¿cuántos sujetos debería haber incluído en el estudio para tener *suficiente* información? Para responder a esta pregunta debemos decidir primero qué magnitud de efecto queremos detectar. Aunque se puede hacer una demostración matemática, el sentido común nos indica que si se quiere detectar un efecto pequeño del tratamiento hará falta más información (una muestra más grande), mientras que para detectar un efecto grande probablemente no hagan falta muchos sujetos. Así que supongamos que estamos dispuestos a considerar como bueno el tratamiento si es capaz de conseguir *en promedio* una disminución de 1 kg de peso en un mes (nótese que decimos en promedio porque siempre va a haber sujetos cuyo peso varíe más y sujetos cuyo peso varíe menos). Si decidimos que nos gustaría tener una probabilidad del 95% de no descartar el tratamiento si realmente adelgaza de promedio 1 kg en un mes, el cálculo de probabilidades nos indica que necesitaríamos 21 sujetos en cada grupo (21 con placebo y 21 con tratamiento)

Ejemplo 3: ensayo de eficacia de la vacuna de Pfizer contra la COVID-19. La figura 6 nos muestra los resultados del ensayo clínico realizado por Pfizer para probar la eficacia de su vacuna contra la COVID. Esta se aplicó en casi 44.000 voluntarios que fueron asignados al azar a dos grupos de 22.000 personas cada uno: un grupo recibió la vacuna y otro recibió el placebo. Ambos grupos eran comparables (misma distribución de sexos, edades, sectores de actividad,...). La figura 6 muestra la incidencia acumulativa diaria (casos de COVID entre los voluntarios de ambos grupos).

Como puede apreciarse, en la primera semana (el recuadro que se muestra ampliado) apenas había diferencia entre ambos grupos (la vacuna aún no había hecho efecto). Pero transcurridos diez días empieza a observarse un efecto claro de la vacuna que se va haciendo más evidente a medida que pasa el tiempo. Mientras en el grupo control (placebo) el número de casos se seguía incrementando sin freno, en el grupo de los vacunados el nivel de contagio era muchísimo menor. Si la vacuna no hubiese tenido efecto, la evolución debería haber sido similar en ambos grupos (como ocurrió la primera semana). Las pruebas estadísticas indicaron que la probabilidad de que esta diferencia se debiera al azar era prácticamente cero; cuando esto ocurre, los estadísticos dicen que el efecto de la vacuna ha resultado *significativo*. Esto quiere decir, en definitiva, que el experimento demostró evidencia suficiente de que el efecto detectado no podía deberse al azar y por tanto la vacuna tenía un efecto real: la incidencia de la enfermedad era *realmente* menor en el grupo vacunado que en el grupo control.

El efecto de la vacuna, además, fue *relevante*; de cada 10.000 no vacunados enfermaron 90; de cada 10.000 vacunados enfermaron 4. Eso significa que, de haberse puesto la vacuna, probablemente 86 de los 90 no vacunados no habrían contraído la enfermedad. Eso da una eficacia de 100 86/90=95.5% con un IC95% de [90.3%, 97.6%] lo que confirmaba la eficacia de la vacuna.

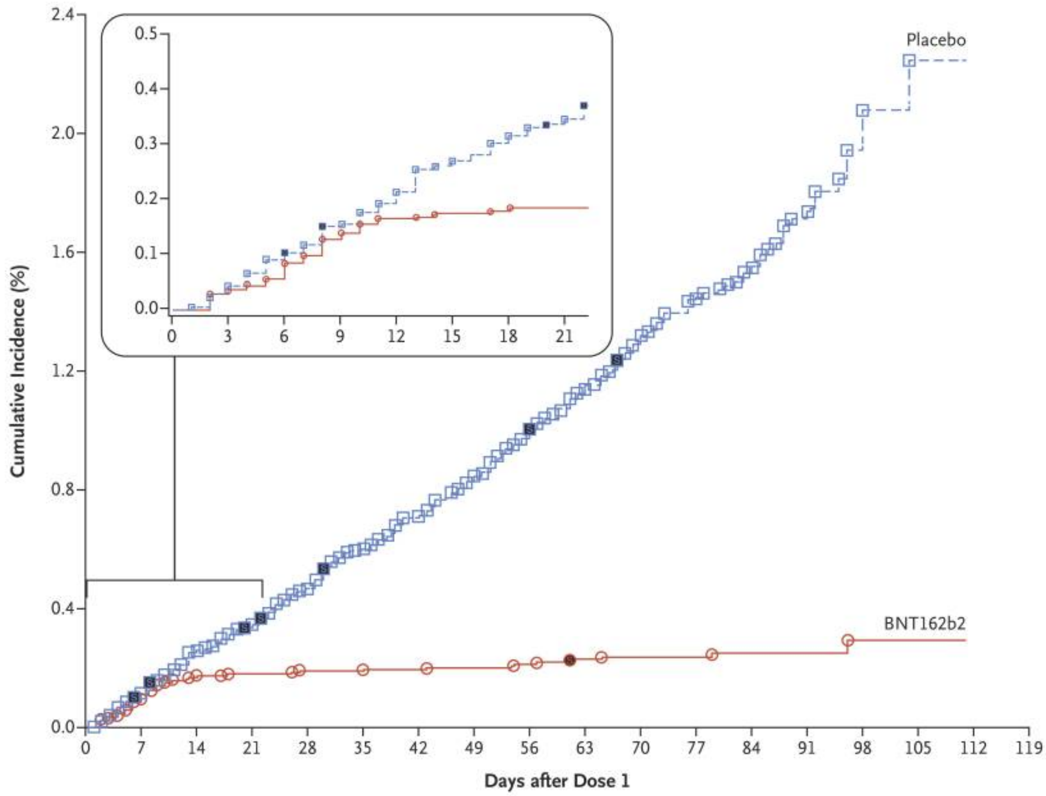

Figura 6. Resultados del ensayo clínico de la vacuna de Pfizer contra la COVID-19.

Ejemplo 4: ¿En qué mes nacen los jugadores de fútbol? La figura 7 siguiente muestra el número de niños (varones) nacidos en España entre 2009 y 2022:

Si asumimos que los nacimientos se reparten por igual a lo largo del año, ello implicaría que cada mes nacerían aproximadamente un $\frac{100}{12} = 8.3\%$ del total del niños del año (con ligeras diferencias debidas a las distintas duraciones de los meses), que es aproximadamente lo que se ve en la figura (quizás con un pequeño exceso de nacimientos en verano respecto a otras estaciones). Sin embargo, si analizamos los meses de nacimiento de los jugadores de fútbol de las ligas españolas de primera y segunda, obtenemos el resultado que se muestra en la figura 8.

Este resultado llama fuertemente la atención, ya que hay un porcentaje muy alto de jugadores que han nacido en los primeros meses del año, y son muchos menos los jugadores nacidos a partir de agosto. ¿Podemos atribuir este resultado a la casualidad? ¿Sucede lo mismo con otras profesiones? La figura 9 es análoga a las anteriores, pero muestra ahora el porcentaje de actores de Hollywood que nace cada mes del año:

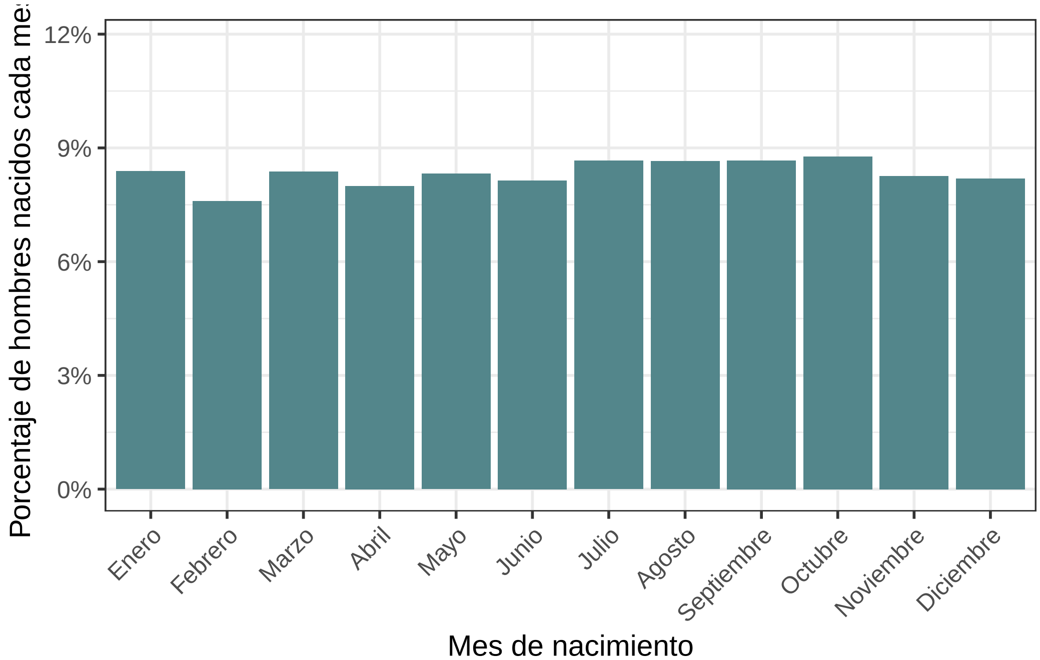

Figura 7. Porcentajes de niños varones nacidos cada mes en España entre 2009 y 2022 (fuente: INE)

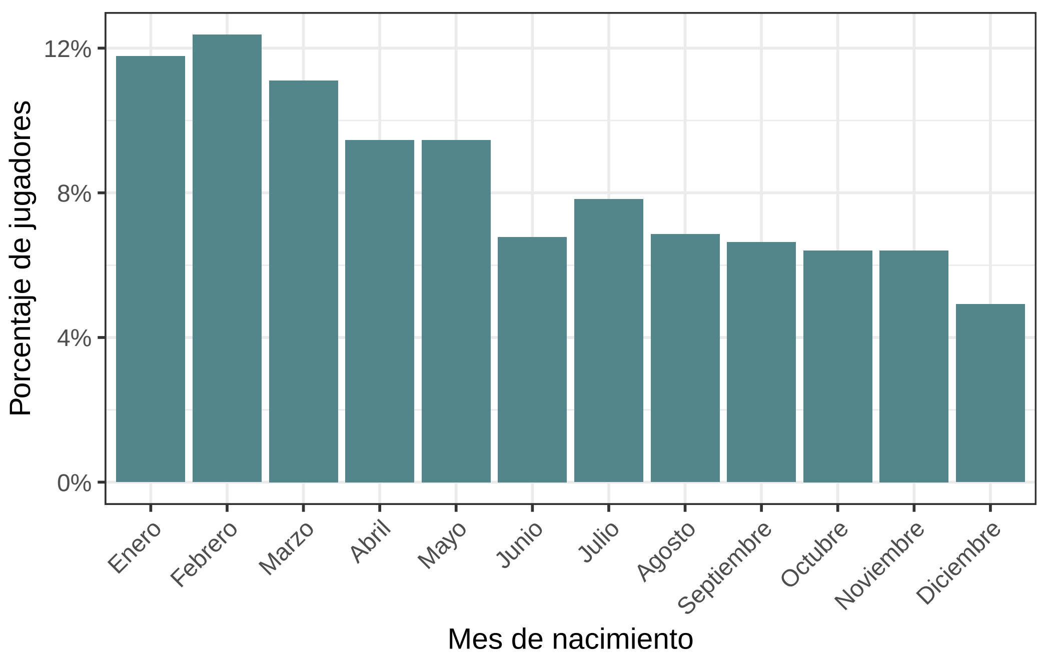

Figura 8. Porcentajes de jugadores de fútbol de la primera y segunda divisiones de España nacidos cada mes (fuente: www.livefutbol.com)

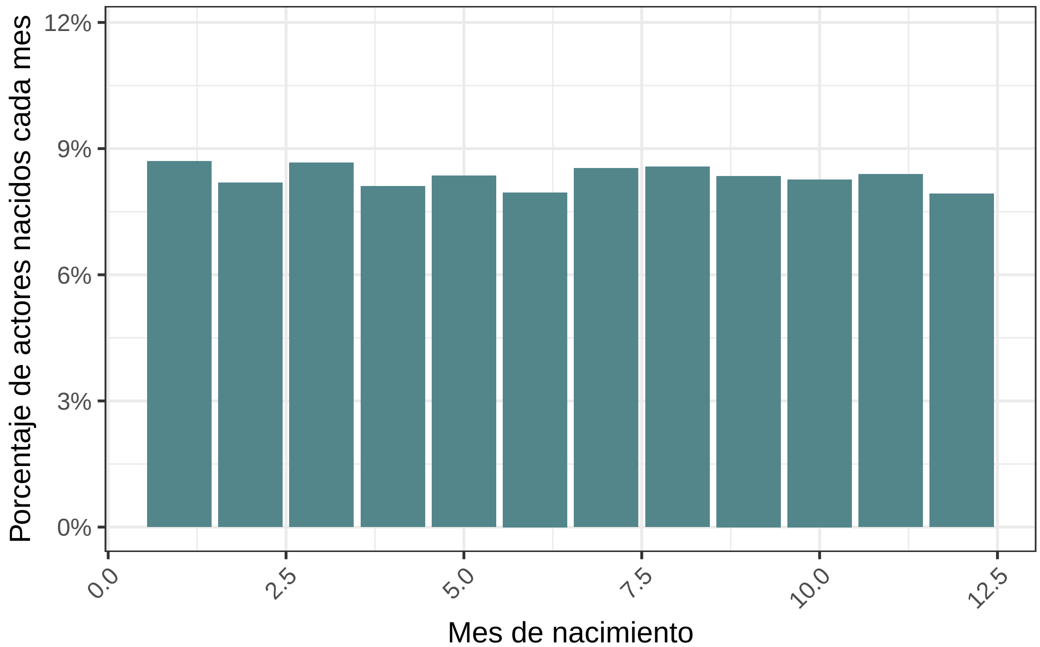

Figura 9. Porcentajes de actores de Hollywood nacidos cada mes (fuente: www.bornglorious.com)

En este caso, la distribución de los meses de nacimiento no parece que se diferencie mucho de la distribución general de nacimientos por mes.

Si utilizamos como hipótesis de partida que los jugadores de fútbol tienen la misma probabilidad de nacer en cualquier mes del año, podemos preguntarnos si esta hipótesis es compatible con los datos, esto es, ¿la acumulación de fechas de nacimiento de los jugadores de fútbol en los primeros meses del año puede haberse producido por casualidad? La aplicación del cálculo de probabilidades nos indica que la probabilidad de que tal cosa ocurra es inferior a una milbillonésima (una entre mil billones), lo que es tanto como decir que es imposible que ésto haya ocurrido por casualidad. Sin embargo, la distibución de meses de nacimiento observada en los actores tiene una no despreciable probabilidad del 16% de haberse producido por puro azar bajo la hipótesis de que tienen la misma probabilidad de nacer en cualquier mes del año.

Así que habrá que preguntarse si hay alguna razón que explique el comportamiento de los meses de nacimiento de los futbolistas. Kiko Llaneras, periodista de El Pais, especializado en artículos de naturaleza estadística, apunta a la posibilidad de que dado que los jugadores de fútbol se forjan en la infancia, en los primeros cursos escolares ocurre que los niños nacidos en los primeros meses del año son mayores que sus compañeros (y a edades tempranas la diferencia de tamaño y habilidades motrices es notable), lo que les hace competir con ventaja en los partidos escolares (y de paso desanimar a los más pequeños de sus cursos). Por ello, muchos de esos niños pasarán a formar parte de equipos juveniles y eventualmente terminarán siendo fichados por los clubs de fútbol profesionales. Es importante señalar que aunque la estadística haya permitido detectar esta curiosa característica de los jugadores de fútbol, la misma carecería de valor si no somos capaces de encontrar una explicación, como la apuntada por Llaneras.

Esta explicación parece en principio razonable. Un poco más de indagación en la web *www.livefutbol.com* nos muestra que el patrón se repite en las ligas de otros países (figura 10)

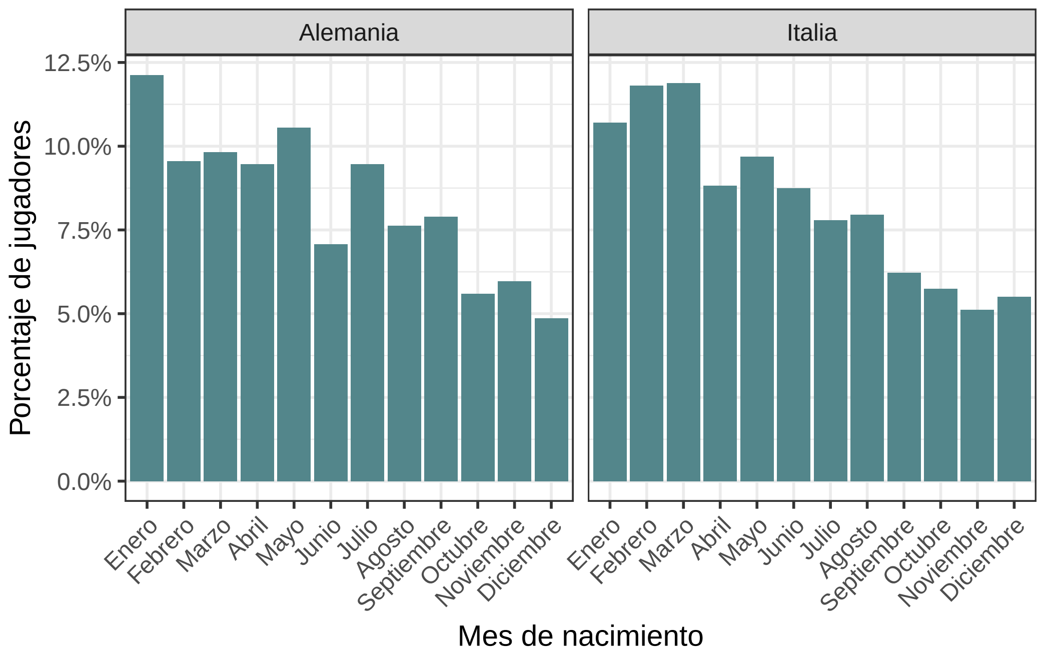

Figura 10. Porcentajes de jugadores de fútbol de las ligas alemana e italiana nacidos cada mes (fuente: www.livefutbol.com)

La repetición del mismo patrón en conjuntos de datos independientes indica que cualquiera que sea la causa de la distribución irregular de fechas de nacimiento de los jugadores de fútbol profesionales, ésta es una propiedad característica de esta población, y que la hace distinta, por ejemplo, de la población de actores.

Ejemplo 5: regresión a la media. Los datos que se muestran en la figura 11 proceden del estudio *"Regression towards mediocrity in hereditary stature"* realizado por Francis Galton en 1886. Representan las alturas (en cm) de padres y madres frente a las estaturas de sus hijos e hijas. Las lineas rectas que hemos dibujado se denominan *rectas de regresión* y representan de manera sencilla la relación entre las estaturas de los hijos y sus progenitores. Ambas rectas son crecientes, reflejando el hecho de que, como puede apreciarse, cuanto más alto es el padre (o la madre) más alto es el hijo (o la hija).

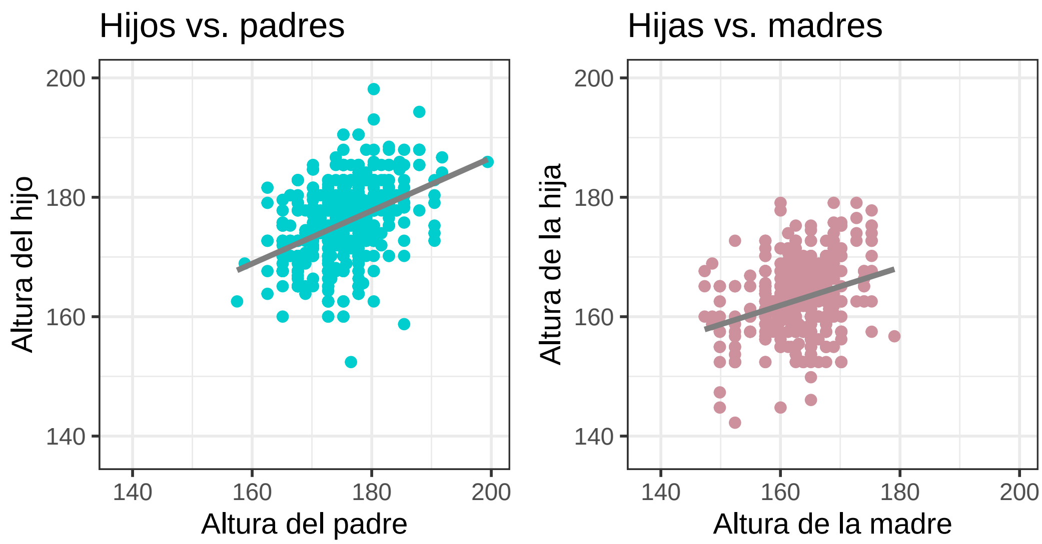

FIGURA 11. Estatura de los hijos frente a la estatura de sus padres en el estudio de Galton de 1886

Si bien ya Gauss había calculado este tipo de rectas en su desarrollo del método de mínimos cuadrados, su uso se popularizó precisamente a partir de este trabajo de Galton. El hecho de denominar *"regresión"* a este modelo obedece a que como consecuencia del mismo, Galton observó que padres (madres) altos tendían a tener hijos (hijas) más bajos que ellos, mientras que padres (madres) bajos tendían a tener hijos (hijas) más altos, lo que significaba de alguna manera que la estatura de los descendientes tendía a volver ("a regresar") a la media. En la tabla 4 hemos separado los padres en tres grupos: los de estatura más alta, los de estatura media y los de estatura más baja. Para cada grupo se muestra la estatura promedio de los padres (madres), así como la estatura de sus hijos (hijas), lo que permite visualizar ese fenómeno de "regresión a la media".

TABLA 4. Tallas medias de los hijos (hijas) según la talla media de sus padres (madres)

Estatura Padre/Madre	**Padres**	**Hijos**	**Madres**	**Hijas**
Baja	167	173	154	161
Estatura media	176	176	163	163
Alta	185	180	171	166

La recta de regresión es un ejemplo de **modelo estadístico**, que es una representación (simplificada) mediante ecuaciones matemáticas de la relación entre variables. Estos modelos se construyen con el objetivo de describir, entender o predecir el comportamiento de una variable en función de otra u otras en base a los datos observados. Estos modelos incluyen además componentes probabilísticos que permite utilizar los valores observados en la muestra para hacer predicciones sobre la población evaluando los márgenes de error probables. Permiten asimismo contrastar hipótesis sobre como es la relación entre las variables involucradas en el modelo.

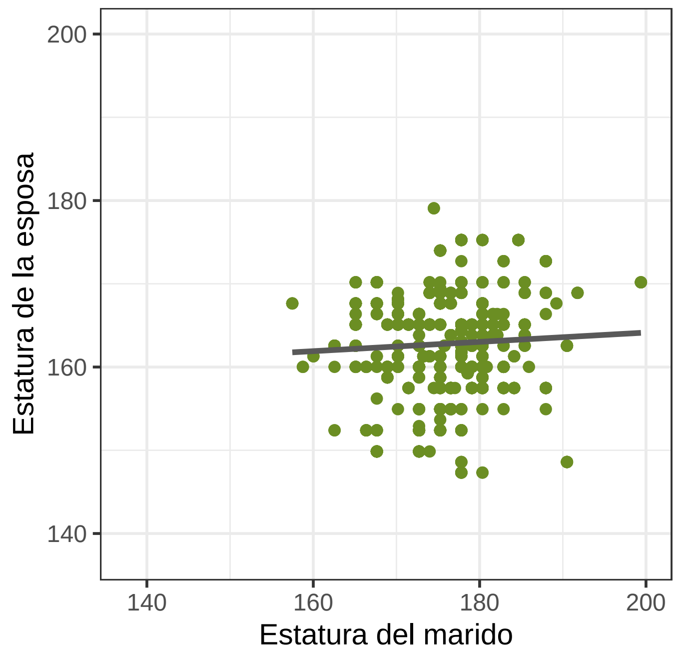

FIGURA 12. Estatura de las esposas frente a la estatura de sus maridos en el estudio de Galton de 1886

A modo de ejemplo, citemos que otro de los objetivos de Galton en su trabajo fue evaluar si hombres altos tendían a casarse con mujeres altas y hombres bajos con mujeres bajas. La figura 12 muestra los valores registrados en su estudio.

En este caso, la recta que relaciona ambas variables aunque está ligeramente inclinada hacia arriba (lo que apuntaría en favor de la hipótesis de Galton) es casi horizontal. El análisis estadístico del ajuste de este modelo indica que si ambas estaturas no estuvieran relacionadas, una inclinación tan leve como la observada sería perfectamente posible por azar (la probabilidad de que esto ocurra es superior al 6.5%, y los científicos se han puesto de acuerdo en que cualquier circunstancia con una probabilidad superior al 5% se puede achacar al efecto del azar). Por tanto los datos son compatibles con la hipótesis de que la estatura no es un criterio de selección a la hora de encontrar pareja. O si lo queremos expresar de otra forma, los datos no muestran evidencia suficiente de que las personas altas tiendan a emparejarse con personas altas y viceversa.

3. El futuro de la Estadística.

Como hemos señalado, la estadística matemática moderna nace de la fusión de la teoría de la probabilidad con el análisis de los datos a través de modelos matemáticos. A partir del desarrollo de los primeros ordenadores en la década de 1950, la informática cobra un papel cada vez más importante en el análisis estadístico, pues permite estudiar conjuntos de datos cada vez más grandes y complejos. A medida que los ordenadores se extienden y se hacen más accesibles se desarrollan nuevas técnicas estadísticas basadas en simulación, se mejoran los algoritmos para el ajuste de modelos estadísticos y se crean diversas herramientas de software que facilitan el análisis de datos: SAS, SPSS, BMDP, Statistica, Systat, ... Entre todas estas herramientas de software destaca el lenguaje R, que se ha convertido actualmente en el estándar para el análisis de datos.

Mientras los estadísticos aprendían a sacar partido de los ordenadores, los expertos en Computación no estaban quietos. En 1957, Frank Rosenblatt inventa el perceptrón, una especie de "neurona artificial" inspirada en el funcionamiento de las neuronas del cerebro humano; el perceptrón recibe información a través de sus "dendritas" y produce un valor de salida que sirve como entrada a otro perceptrón. La conexión de diversos perceptrones entre sí forma lo que se conoce como "red neuronal" que, adecuadamente programada, es capaz de realizar tareas de clasificación. Estas redes neuronales, junto con otros algoritmos, continuaron su desarrollo durante los años 60 y 70. En los años 80 las redes neuronales ya empezaban a ser capaces de identificar patrones complejos y a partir de ese momento en el mundo de la computación se empieza a hablar de "aprendizaje automático" o "machine learning". Estas técnicas se van volviendo cada vez más sofisticadas y empiezan a realizar tareas que tradicionalmente se asociaban al campo de la estadística; por ejemplo, los métodos estadísticos de regresión utilizan los valores de una serie de variables *explicativas* para predecir los valores de una variable respuesta. Los algoritmos de *machine learning* empezaron a ser capaces de realizar esta tarea "aprendiendo" de los datos. Así, en 1998 google empezó a utilizar técnicas de machine learning para "predecir" en qué lugar de la web se encontraba la información que buscaba un usuario. Los nuevos algoritmos eran capaces de utilizar grandes bases de datos (lo que hoy llamamos *big data*) para hacer predicciones: cuál sería el comportamiento de la bolsa, cuáles sería la demanda de productos en función de los perfiles de los usuarios, ... En el año 2010 nacen las técnicas de *deep learning* (aprendizaje profundo) que permiten que las máquinas sean capaces de reconocer el contenido de imágenes, de leer texto escrito a mano, de entender el lenguaje natural, de realizar traducciones, ... Y así hasta el momento actual, en que los algoritmos de *"Inteligencia artificial"*, evolución de todos los anteriores, empiezan a formar parte de nuestro día a día: chatGPT, Llama, Bard-Gemini, Claude, Perplexity, ...

Todos estos métodos se van integrando en el cajón de herramientas de la Estadística, aunque probablemente un experto en Computación diría que son los métodos estadísticos los que se van integrando en el cajón del *machine learning*. Sea como sea, lo cierto es que se se ha producido -se está produciendo- la convergencia entre ambos mundos. Las posibilidades que se abren para el futuro son múltiples. Al escribir el texto de esta charla he hecho el ejercicio de preguntarle a varias inteligencias artificiales (a través del portal https://chat.lmsys.org/) cuál será el futuro de la estadística en los próximos años. Las respuestas de las distintas IA fueron bastante coincidentes y se muestran en los párrafos siguientes:

El futuro de la estadística es prometedor y se espera que juegue un papel aún más crítico en diversas áreas del conocimiento y la actividad humana. A continuación se mencionan algunos aspectos en los que la estadística probablemente seguirá evolucionando y desempeñando un papel importante:

1. **Big Data y análisis de datos masivos**: Con la creciente cantidad de datos generados por dispositivos electrónicos, sensores y sistemas en línea, la estadística es esencial para analizar, interpretar y extraer información útil de estos grandes volúmenes de datos. En particular, se espera que se desarrollen nuevos métodos estadísticos que se adapten a la escala de la nube y permitan analizar datos en tiempo real. Esto será crucial para áreas como la economía, la finanzas, la salud pública y la gestión de riesgos.

2. **Inteligencia Artificial y aprendizaje automático**: La estadística es fundamental en el desarrollo y la mejora de algoritmos de aprendizaje automático, que son el corazón de muchos sistemas de inteligencia artificial. La comprensión de los principios estadísticos ayuda en la creación de modelos predictivos más precisos y eficientes. Se espera un mayor uso de estos modelos para automatizar tareas, identificar patrones complejos y generar predicciones más precisas.

3. **Ciencia de datos**: La estadística es una de las disciplinas centrales en el campo emergente de la ciencia de datos, entendida ésta como el estudio de los procedimientos que permiten a gobiernos, empresas y organizaciones convertir sus datos en conocimiento accionable que haga posible tomar decisiones informadas y basadas en evidencia; por ello, la demanda de habilidades estadísticas aumentará.

4. **Investigación en salud**: En áreas como la epidemiología, la bioestadística y la investigación en salud pública, la estadística seguirá siendo esencial para la investigación de enfermedades, el desarrollo de tratamientos, el diseño de estudios, el análisis de datos clínicos y la toma de decisiones basada en evidencia.

5. **Educación y formación en estadística**: A medida que la importancia de la estadística sigue creciendo, también lo hace la necesidad de educación y formación estadística. Se espera que los programas educativos sigan evolucionando para satisfacer esta demanda, y que se fomente una mayor alfabetización estadística que permita a los estudiantes comprender y analizar datos de manera efectiva.

6. **Metodologías estadísticas**: Las técnicas y metodologías estadísticas continuarán desarrollándose y refinándose. El avance en el poder computacional y los algoritmos permitirá explorar técnicas más complejas y sofisticadas que pueden manejar mejor las peculiaridades de los datos modernos, como la alta dimensionalidad, los datos geolocalizados y la estructura no estándar que se presenta en datos como los procedentes de redes sociales, de sensores de todo tipo y de la Internet de las Cosas.

7. **Transparencia y reproducibilidad**: Habrá un énfasis continuo en la transparencia estadística y la reproducibilidad de los resultados de investigación, lo que puede llevar a prácticas más rigurosas en el registro de análisis de datos y en la publicación de conjuntos de datos y metodologías utilizadas.

8. **Ética y privacidad**: A medida que la capacidad de recolectar y analizar datos crece, también lo hace la preocupación por la privacidad y la ética. La estadística estará en primera línea para desarrollar métodos que protejan la privacidad de los individuos mientras se extrae conocimiento de los datos. Se deberá garantizar que los datos se utilicen de manera responsable y que se protejan los derechos de las personas.

9. **Colaboración interdisciplinaria**: La estadística seguirá siendo un campo altamente interdisciplinario, colaborando con expertos en otras áreas para resolver problemas complejos y contribuir al progreso en campos tan diversos como las ciencias sociales, las finanzas, la ingeniería y más.

10. **Integración de la estadística con la interpretación**: La interpretación de los resultados estadísticos es cada vez más importante en la toma de decisiones.

Se deberá garantizar que los resultados estadísticos sean interpretados de manera precisa y que se tomen decisiones informadas.

11. **Enfoque en la visualización**: La visualización de datos se volverá aún más importante para comunicar información compleja de manera clara y accesible a una audiencia más amplia. Se espera un auge de herramientas y técnicas innovadoras para crear visualizaciones interactivas e inmersivas que permitan una mejor comprensión de los datos.

12. **Democratización de la estadística**: Se facilitará el acceso a las herramientas y técnicas estadísticas para que un público más amplio pueda entenderlas y utilizarlas.

En resumen, el futuro de la estadística estará marcado por la innovación tecnológica, la colaboración interdisciplinaria y la búsqueda de soluciones a los grandes desafíos de la sociedad. Se espera que esta disciplina siga jugando un papel fundamental en la toma de decisiones informadas y en la construcción de un futuro mejor."

Tras leer las respuestas de las IA, poco más se puede añadir. Si acaso hay algo que llama la atención es la insistencia en la formación estadística de las personas, aún cuando es muy probable que en los próximos años muchos análisis estadísticos van a ser realizados directamente por una IA sin intervención de un estadístico (humano) profesional (véase en el apéndice las respuestas dadas, respectivamente por Bard/Gemini, la IA de Google, y chatGPT de openAI al problema de decidir si el tratamiento para adelgazar que vimos en el ejemplo era o no efectivo). Ahora bien, como en última instancia las decisiones habrán de ser tomadas por personas, éstas deberán tener la formación y conocimientos suficientes para ser capaces de entender y valorar críticamente los análisis realizados por la IA para que sus decisiones sean lo más justas y adecuadas posible.

4. Apéndice

Tanto a la IA de Google (bard.google.com) como a la de openAI (chatGPT) se les ha planteado la cuestión siguiente:

Para responder a la siguiente cuestión debes ponerte en el papel de un estadístico profesional. Se dispone de datos de 10 personas a las que se les ha aplicado un tratamiento para bajar de peso. La variación de peso resultante en estas personas fue: 0.127, -0.628, -1.521, -0.856, 0.587, -1.976, 0.194, -1.365, 0.213, -1.116. Los datos se expresan en kilogramos. Asimismo se utilizaron otras diez personas como control (se les dio un placebo en lugar de tratamiento). En estas personas, la variación de peso fue: -0.641, 0.251, -1.401, -0.119, -1.286, -0.22, 1.726, 0.277, 0.042 y 0.546 kilogramos, respectivamente. ¿Como decidirías si el tratamiento fue efectivo para perder peso?

Respuesta de Google Bard:

Análisis estadístico para determinar la efectividad del tratamiento

1. Planteamiento del problema:

Se desea determinar si un tratamiento para bajar de peso es efectivo. Para ello, se cuenta con datos de 10 personas que recibieron el tratamiento y 10 personas que recibieron un placebo (grupo control). Se observa la variación de peso en cada persona durante un periodo de tiempo.

2. Hipótesis:

- Hipótesis nula (H0): El tratamiento no tiene ningún efecto sobre la pérdida de peso. La variación de peso en el grupo de tratamiento y en el grupo control será similar.

- Hipótesis alternativa (H1): El tratamiento tiene un efecto positivo sobre la pérdida de peso. La variación de peso en el grupo de tratamiento será significativamente mayor que en el grupo control.

3. Prueba estadística:

Para comparar las medias de los dos grupos (tratamiento y control) se puede realizar una prueba t para dos muestras independientes. Esta prueba nos permite determinar si la diferencia entre las medias es significativa o no.

4. Análisis de datos:

Media del grupo de tratamiento: -0.634 kg Media del grupo control: -0.2849 kg Desviación estándar del grupo de tratamiento: 0.874 kg Desviación estándar del grupo control: 1.431 kg Valor p de la prueba t: 0.26

5. Interpretación de resultados:

El valor p de la prueba t es de 0.26. Este valor es mayor que el nivel de significancia preestablecido (generalmente 0.05). Por lo tanto, no podemos rechazar la hipótesis nula. Esto significa que no hay evidencia suficiente para concluir que el tratamiento tiene un efecto significativo sobre la pérdida de peso.

6. Conclusiones:

Con base en los datos disponibles, no se puede concluir que el tratamiento sea efectivo para perder peso. Se necesitan más estudios con un mayor número de participantes para confirmar o refutar esta conclusión.

7. Limitaciones del estudio:

Es importante tener en cuenta que este estudio tiene algunas limitaciones:

- El tamaño de la muestra es pequeño (solo 10 personas por grupo).

- No se controlaron otras variables que podrían afectar la pérdida de peso, como la dieta o la actividad física.

- El estudio no se realizó de forma ciega, lo que podría introducir sesgos en los resultados.

8. Recomendaciones:

Para obtener resultados más concluyentes, se recomienda realizar estudios con un mayor número de participantes, controlar otras variables que podrían afectar la pérdida de peso y realizar el estudio de forma ciega.

En resumen, el análisis estadístico no proporciona evidencia suficiente para concluir que el tratamiento es efectivo para perder peso. Se necesitan más estudios para confirmar o refutar esta conclusión.

Respuesta de chatGPT (chat.openai.com):

Para evaluar la efectividad del tratamiento para bajar de peso, realizaré un análisis comparativo entre el grupo de tratamiento y el grupo de control. Utilizaré estadísticas descriptivas y pruebas de hipótesis para llegar a una conclusión.

Primero, calcularé las estadísticas descriptivas, como la media y la desviación estándar, para cada grupo. Esto proporcionará una visión general de la tendencia central y la dispersión de los datos.

Grupo de tratamiento: Media = -0.634 Desviación estándar = 0.874

Grupo de control: Media = -0.2849 Desviación estándar = 1.431

Ahora, compararé las medias de ambos grupos. Para determinar si hay una diferencia significativa entre los grupos, puedo realizar una prueba t de Student (prueba t) para muestras independientes. La hipótesis nula (H0) sería que no hay diferencia significativa en la variación de peso entre los dos grupos, mientras que la hipótesis alternativa (H1) sería que hay una diferencia significativa.

Realizar la prueba t y obtener el valor p asociado. Si el valor p es menor que un nivel de significancia previamente establecido (por ejemplo, 0.05), rechazaríamos la hipótesis nula y concluiríamos que hay evidencia suficiente para decir que el tratamiento tuvo un efecto significativo en la pérdida de peso.

En resumen, decidiría si el tratamiento fue efectivo para perder peso mediante un análisis estadístico comparativo, utilizando pruebas de hipótesis para determinar si hay una diferencia significativa entre el grupo de tratamiento y el grupo de control en términos de variación de peso.

Comentario. Llama la atención el detalle y precisión de la respuesta de Bard, señalando la conclusión correcta e indicando las limitaciones del experimento. La respuesta de chatGPT es bastante más pobre en este sentido, y aunque es correcta no da un resultado o conclusión final. Es posible que esto pueda achacarse a que aquí hemos usado la versión 3.5 de chatGPT (es la que permite el uso gratuito) y es muy probable que la versión 4 hubiese mostrado una respuesta similar a la de Google.

Referencias

1. Basulto, J., Camúñez, J.A. (2007) El problema de los dados del caballero de Méré: soluciones publicadas en el siglo XVII SUMA vol 56, pp 43-54. Disponible en http://materias.df.uba.ar/estadisticaa2019v/files/2019/02/El_caballero_de_Mere.pdf

2. Fernández, S. (2006) El azar y sus problemas. SUMA, 51, pp 99-105. Disponible en https://revistasuma.fespm.es/sites/revistasuma.fespm.es/IMG/pdf/51/099-105.pdf

3. Fernández, S. (2007) Los inicios de la teoría de la probabilidad. SUMA vol. 55, pp 7-20. Disponible en https://revistasuma.fespm.es/sites/revistasuma.fespm.es/IMG/pdf/55/007-020.pdf

4. Galton, F. (1886). «Regression towards mediocrity in hereditary stature». The Journal of the Anthropological Institute of Great Britain and Ireland 15: 246-263. JSTOR 2841583. doi:10.2307/2841583. Disponible en https://www.jstor.org/stable/2841583

5. Ibarrola, P. (2006) Gauss y la Estadística. Conferencia pronunciada en el curso 2005-2006 en la Facultad de Matemáticas y Estadística de la Universidad Politécnica de Valencia. Disponible en https://fme.upc.edu/ca/arxius/butlleti-digital/gauss/060215_conferencia_ibarrola.pdf

6. Llaneras, K. (2022) Piensa Claro: ocho reglas para descifrar el mundo y tener éxito en la era de los datos. Ed. Debate

7. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, Perez JL, Pérez Marc G, Moreira ED, Zerbini C, Bailey R, Swanson KA, Roychoudhury S, Koury K, Li P, Kalina WV, Cooper D, Frenck RW Jr, Hammitt LL, Türeci Ö, Nell H, Schaefer A, Ünal S, Tresnan DB, Mather S, Dormitzer PR, Şahin U, Jansen KU, Gruber WC; C4591001 Clinical Trial Group. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. N Engl J Med. 2020 Dec 31;383(27):2603-2615. doi: 10.1056/NEJMoa2034577. Epub 2020 Dec 10. PMID: 33301246; PMCID: PMC7745181. Disponible en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7745181/

8. Schönberger, V.M., Cukier, K. (2013). Big Data: La revolución de los datos masivos. Turner Publicaciones, Madrid.

9. Stigler, S.M. (1990) The history of Statistics. The measurement of uncertainty before 1900. Harvard University Press.