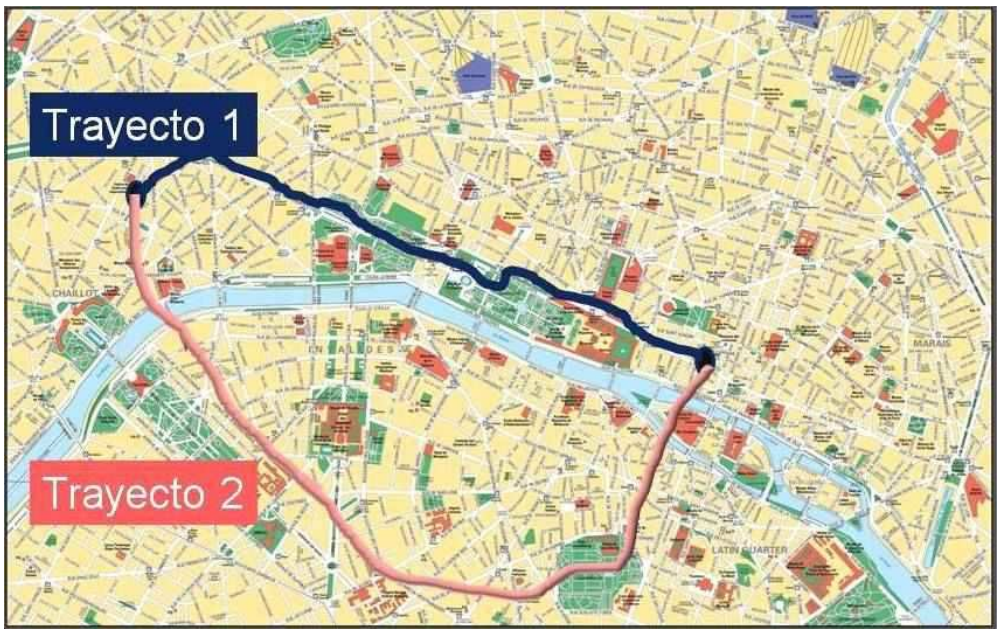
Una fábula significativa

La necesidad de una decisión

El Sr. Traveller regresa en automóvil diariamente desde su centro de trabajo a su domicilio en París. En principio tiene dos opciones: un trayecto más largo pero que se extiende a lo largo de un espacio poco transitado, y otro más corto, pero frecuentemente más abarrotado por el tráfico.



En principio, Traveller quiere demorar lo menos posible en arribar a su casa debido a que su madre está enferma y procura estar con ella la mayor parte de su tiempo libre; de modo que necesita conocer el tiempo que insume cada una de las dos opciones. Obviamente, se trata del tiempo promedio, pues está claro que un mismo trayecto producirá diferentes registros en dependencia de las condiciones climáticas, de que se produzcan o no atascos, del día de la semana y de otros diversos elementos azarosos que influyen en la velocidad con la que se cumplen cada uno de los recorridos.

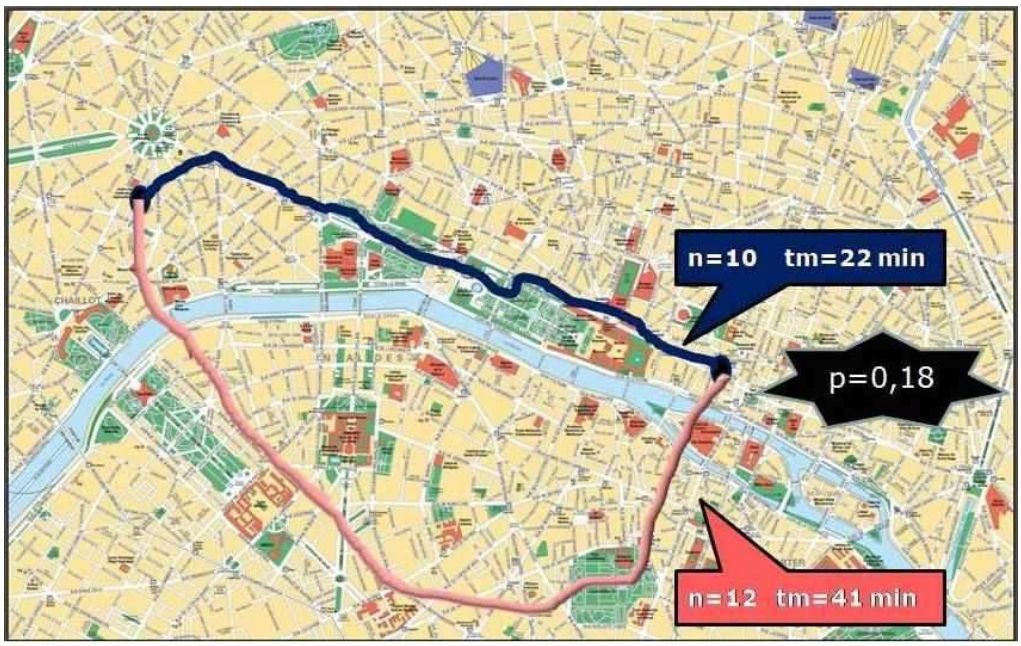
Observaciones y pruebas de significación

Traveller decide hacer algunas observaciones a lo largo de un mes. Concretamente, elige de manera aleatoria, cada día laboral de ese mes, uno de los dos trayectos. Registra el resultado de $n_1 = 10$ y $n_2 = 12$ viajes por una y otra vía respectivamente y, al concluir el mes, obtiene los datos que se recogen en la Tabla 1:

Tabla 1. Minutos invertidos para cumplir el recorrido en n_1=10 ocasiones por en el Trayecto 1 y n_2=12 ocasiones por el Trayecto 2

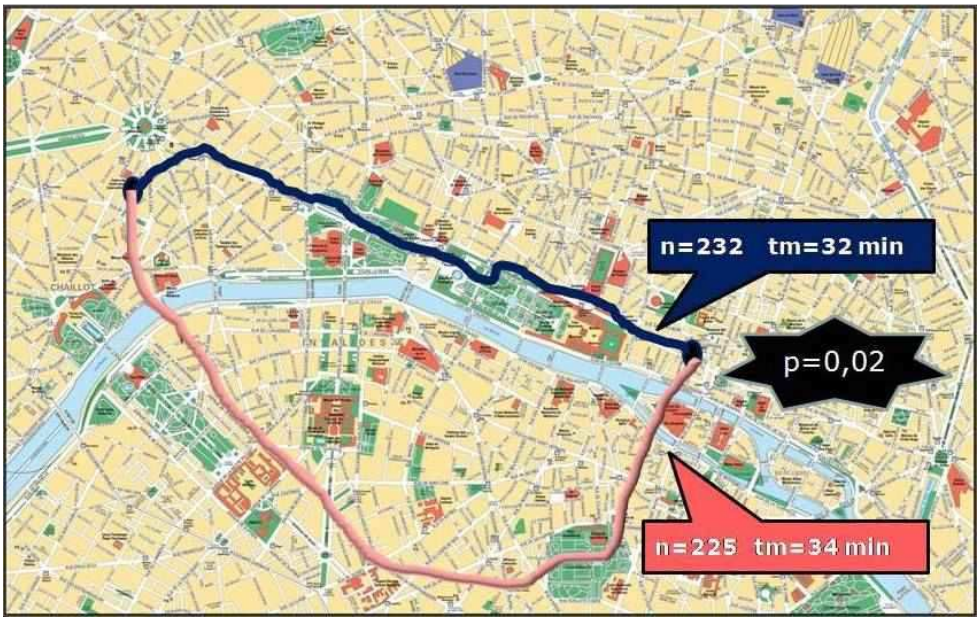
Tiempos	1	2	3	4	5	6	7	8	9	10	11	12
Trayecto1	12	7	10	11	10	12	9	54	30	62		
Trayecto2	97	15	13	91	8	15	14	11	17	85	10	115

Es fácil corroborar que el tiempo medio invertido en el primer trayecto asciende a 22 minutos, en tanto que el segundo insume casi el doble: 41 minutos. El trayecto más corto resulta ser también el que menos tiempo exige y la diferencia es muy apreciable. Pero una prueba convencional de comparación de medias (t de Student) arroja que la p asociada a esa diferencia de 19 minutos es p=0,18.



Puesto que no puede descartarse el azar como explicación, Traveller considera que no puede concluir nada. Consecuentemente, en lugar de tomar una decisión acerca de cuál es la vía más apropiada, decide realizar un número mucho mayor de viajes por uno y otro camino para adoptar entonces su decisión.

Concretamente, reproduce el proceso, pero ahora en n_1=232 y n_2=225 ocasiones respectivamente. Los resultados de esas experiencias se recogen en el archivo **paris.xls**. Es fácil corroborar que el tiempo medio sigue siendo menor para el primero de los trayectos que para el segundo (32 y 34 minutos respectivamente). La misma prueba t de Student arroja ahora un valor p=0,02, muy por debajo del umbral convencional $\alpha = 0{,}05$, de modo que puede declararse que la diferencia de 2 minutos es estadísticamente significativa.



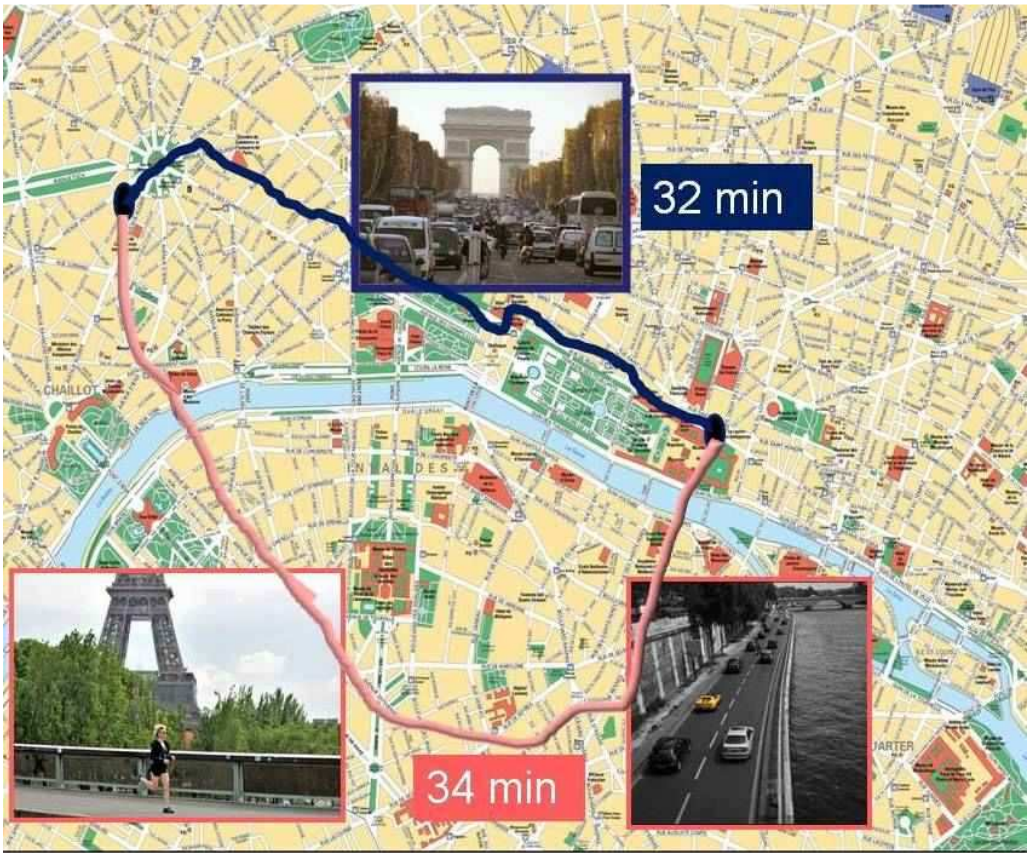
La diferencia hallada resultó ser esta vez estadísticamente significativa. A la luz de estos datos, teóricamente, Traveller debería elegir el primer trayecto para sus próximos desplazamientos.

La decisión

Para interpretar adecuadamente los resultados y tomar la decisión más adecuada, Traveller revisa varios artículos donde se emplean pruebas de significación y registra que, una y otra vez, en situaciones similares, se dice que "la diferencia encontrada es significativa" o que "el efecto es significativo". De acuerdo con eso, podría afirmarse que "el primer trayecto es significativamente más corto que el segundo".

Sin embargo, sospecha que, así dicho, se estaría cometiendo un abuso de lenguaje, ya que lo que realmente ocurre no es que "la diferencia sea significativa" sino que "la diferencia es *estadísticamente* significativa". Traveller trata de comprender qué distingue a una afirmación de la otra; es decir, qué agrega o que quita el adverbio "estadísticamente". Le resulta evidente que en el primer caso, lo que se está sugiriendo implícitamente es que la diferencia en sí (en su caso, los dos minutos) es trascendente, mientras que la segunda formulación aclara que dicha diferencia no necesariamente tiene algún significado independiente de la estadística. Esta segunda –que es la correcta- no informa sobre la importancia de la diferencia, sino sobre sino el resultado de la prueba de significación estadística realizada. Siendo así, intuye que regirse mecánicamente según esa regla, acaso no sea lo más conveniente.

En este punto, Traveller procede a examinar qué significan realmente en sus circunstancias esos dos minutos. Considera que tal diferencia carece de importancia práctica. Resulta para él evidente que emplear dos minutos adicionales es un percance mucho menor que el que sufrirán su sistema nervioso y sus coronarias como consecuencia de hacer un recorrido más corto en medio de avenidas atestadas por el tránsito. Su larga experiencia con el trayecto más largole indica, además, que el segundo recorrido le permite realizar un viaje mucho más placentero: podrá disfrutar sin tensiones el paisaje mientras observa a los parisinos que trotan por los parques aledaños.



Tras esas consideraciones, toma la decisión final que le parece más sensata: elige el segundo trayecto, opción opuesta a la que indica su prueba de significación. En síntesis, las pruebas de significación no le fueron de ninguna utilidad, ni cuando no obtuvo significación ni cuando la obtuvo. En vista de este frustrado intento por emplearlas para resolver un problema concreto, Traveller profundiza en su reflexión sobre la metodología en cuestión.

Una sospecha a partir de la intuición

Su sentido común indica a Traveller que la duración media real (la que se obtendría si se hicieran infinitos viajes) para el Trayecto 1 (llamémosle Δ_1) no puede ser exactamente igual a la del Trayecto 2 (que se denotará por Δ_2). La diferencia $\Delta = \Delta_2 - \Delta_1$ no será exactamente igual a cero por la sencilla razón de que los trayectos son diferentes; no existe razón alguna para pensar que pudiera ser válida tan inverosímil posibilidad, del mismo modo que sería absurdo creer que dos árboles de un bosque puedan ser idénticos entre sí. Intuye también que en la medida que los tamaños de muestra sean mayores (que se hagan más experiencias usando ambos recorridos) más verosímil será que la estimación de la diferencia se acerque al verdadero Δ. De modo que la diferencia muestral de 19 minutos inicialmente obtenida constituye una estimación poco confiable de Δ, ya que se obtuvo sobre la base de muy pocas observaciones; en cambio, la segunda, de 2 minutos, resulta mucho más digna de crédito por haberse conseguido usando cientos de observaciones, y es por tanto mucho menos susceptible a las veleidades del azar.

Finalmente, intuye que con el incremento de los tamaños muestrales, será más probable que la prueba de significación "detecte" la verdad: que tal diferencia no es nula. En términos del valor empleado para declarar o no significación, esto equivale a decir que cuanto mayores sean los tamaños de muestra, menor será el valor p que habrá de comprarse con el umbral fijo que se haya fijado, cualquiera sea éste (independientemente de que ya se sabe que, convencionalmente, suele elegirse $\alpha = 0{,}05$)

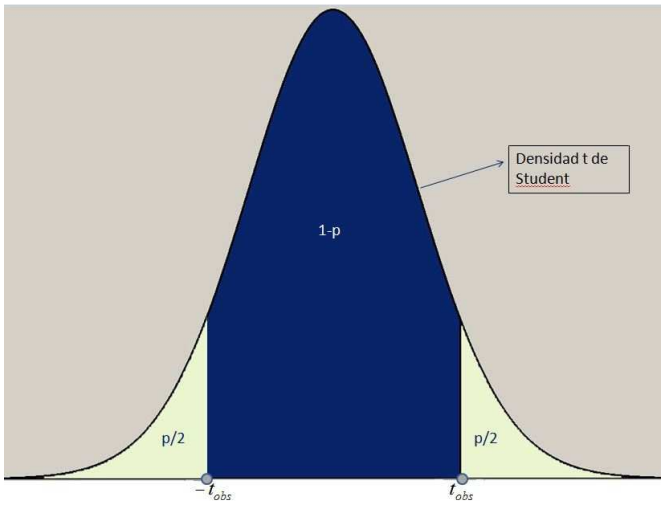
Matemática y sentido común

En ese punto, Traveller explora la naturaleza matemática de dicha prueba, esperando que ésta refrende lo que le dice su intuición. Su libro de estadística inferencial favorito le permite reparar en que la prueba t de Student funciona del modo siguiente:

Supongamos que tenemos n_1 observaciones en la muestra 1 y n_2 observaciones en la muestra 2, así como que llamamos $x_1, \cdots, x_{n_1}$ y $y_1, \cdots, y_{n_2}$ a las observaciones respectivamente obtenidas. Llamando $\bar{\boldsymbol{x}}$ y $\bar{\boldsymbol{y}}$ a las medias de las muestras respectivas, se obtiene la diferencia muestral de medias $\boldsymbol{d} = \bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}$, que es una estimación de Δ. La valoración de la hipótesis nula que afirma que Δ=0 se realiza calculando el siguiente estadístico:

$$t_{obs} = \frac{d}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde s_1 es la desviación estándar asociada a la primera muestra, s_2 la que corresponde a la segunda y s se calcula mediante $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$. Para obtener el valor p, se calcula el doble del área a la derecha de t_{obs} bajo la curva de densidad asociada a una variable que se distribuye t de Student con n grados de libertad ($n = n_1 + n_2 - 2$). En términos gráficos, esto es:



Es ahora obvio que cuanto mayor sea t_{obs}, menor será el valor de p. Pero a su vez, también lo es que t_{obs} crecerá en la medida que sean mayores los tamaños muestrales n_1 y n_2. Dicho de otro modo: salvo en la insólita situación en que los trayectos tengan exactamente igual duración, para obtener significación sólo hará falta haber hecho un número suficientemente grande de observaciones. Queda confirmado que el "sentido común" de Traveller no resulta traicionado por la matemática en lo que concierne a la prueba de significación.

Llegado al punto, en que corrobora que éstas no le sirven para nada a los efectos de tomar una decisión, se pregunta cuán confiable son las estimaciones de Δ (19 y 2 minutos respectivamente) las cuales sí fueron incluidas entre los elementos tenidos en cuenta para decidir.

Nuevamente, apela a su libro de estadística inferencial. Allí se explica que la forma convencional de resolver ese problema es empleando intervalos de confianza; un intervalo dentro del cual uno puede estar "confiado" que se incluirá el parámetro, Δ en este caso. Para el problema que le ocupa, los límites para un intervalo de confianza al $(1-\alpha)*100\%$, donde α es un número entre 0 y 1 (usualmente igual a 0,05) vienen dados por las fórmulas siguientes:

$$\left(d - t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \; ; \; d + t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

donde t_α es el percentil $(1-\frac{\alpha}{2})100\%$ de la distribución t de Student con n1+n2-2 grados de libertad.

El hecho de que aquello que se resta de y se suma a d para tal construcción es una expresión que tiene los tamaños muestrales en el denominador, confirma la intuición de Traveller, según la cual cuanto mayores sean éstos, mejor será la aproximación que ofrece la estimación d al conocimiento real de Δ. Al realizar el cálculo con los datos de la primera experiencia (usando $\alpha = 0{,}05$ y la información de la Tabla 1), se obtiene el intervalo: $(-10{,}1\,;\,48{,}1)$. Esto quiere decir que podemos estar confiados en que la diferencia a favor del Trayecto 2 no superaría los 48,1 minutos, pero también que dicho trayecto pudiera suponer una inversión de hasta 10,1 minutos menos que si se escoge el Trayecto 1.

Cuando se hace lo propio con los datos que figuran en parix.xls (el caso en que n1=232 y n2=225), se obtiene un intervalo de confianza considerablemente más estrecho. $(0{,}1\,;\,3{,}9)$. Esto significa que Traveller puede estar altamente confiado de que siguiendo el segundo Trayecto no producirá, como promedio, un aumento del tiempo invertido superior a 3,9 minutos, caso extremo que, a los efectos prácticos, tampoco tendría mayor importancia.

La perspectiva de los intervalos de confianza sí le resulta útil: le permite ratificar como correcta y oportuna su decisión.

Traveller encara un problema más serio

El problema de los trayectos es relativamente baladí para Traveller. Así es, sobre todo si lo compara con una decisión muchísimo más trascendente que espera por él. En efecto, su madre está enferma y él ha

de decidir qué tratamiento se le aplicará. Él tiene que sufragar los costos de aquel que se elija, la decisión le concierne directamente, ya que es él quien a la postre habrá de adoptarla.

Se le informa que hay dos tratamientos disponibles: uno convencional (**C**) y otro novedoso (**N**). El dilema se resolverá cuando Traveller consiga aquilatar sus méritos relativos. Según le informan los médicos, algo que él corrobora a través de Internet, se han hecho dos ensayos clínicos controlados para la grave dolencia que aqueja a su progenitora. En ambos se evaluó la sobrevivencia transcurrido un año desde el inicio del tratamiento.

El primero se realizó en un hospital comarcal y arroja que la tasa de mortalidad correspondiente a **N** es mucho menor que la registrada cuando se aplica **C**. Los resultados concretos figuran en la Tabla 2.

Tabla 2. Muerte y sobrevivencia tras un año desde el comienzo de dos tratamientos en un pequeño ensayo clínico llevado adelante en un hospital

	Tratamiento **N**	Tratamiento **C**
Mueren	6	9
Sobreviven	12	9
Total	**18**	**18**

Las tasas fueron 33% (6/18) para el tratamiento **N** y 50% (9/18) para el tratamiento **C**. Consecuentemente, el tratamiento **N** parece más atractivo. Sin embargo, al comparar dichas tasas mediante una prueba de Ji-cuadrado, se obtiene una p = 0,3. El artículo que da cuenta de este estudio concluye que “no hay evidencias que permitan afirmar que las tasas sean diferentes”.

El segundo estudio ha sido llevado adelante con la financiación de una importante transnacional del medicamento (la cual comercializa el producto en que se basa el primer tratamiento), de manera que se consiguió realizar un enorme estudio multicéntrico que involucró a más de 10 mil pacientes a lo largo de varios años. Los tamaños de las muestras fueron 5000 (Tratamiento **N**) y 5180 (Tratamiento **C**), y dicho estudio arroja que el 32% e las personas murieron después de recibir el tratamiento **C** ($t_C = 0{,}32$), mientras que esta cifra fue del 30% para quienes recibieron el **N** ($t_N = 0{,}30$). La Tabla 3 muestra los resultados dispuestos en una tabla convencional de 2X2:

Tabla 3. Muerte y sobrevivencia transcurrido un año luego del comienzo de dos tratamientos en un ensayo clínico controlado multicéntrico

	Tratamiento **N**	Tratamiento **C**
Mueren	1500	1658
Sobreviven	3500	3522
Total	**5000**	**5180**

Realizada la misma prueba de Ji-Cuadrado arriba mencionada, se obtiene que un valor de p igual a 0,008. Siendo ese número mucho menor que 0,05, el artículo correspondiente afirma que la diferencia $d = t_C - t_N = 0{,}02$ es estadísticamente muy significativa y agrega, no sin razón, que el tratamiento **N** puede considerarse más efectivo que el **C** a los efectos de la supervivenvia. A la luz de estos resultados

y siguiendo el mismo razonamiento que realizó cuando estudiaba el tema de los trayectos por las calles parisinas, Traveller llega a la conclusión de que las tasas reales de mortalidad no serán exactamente iguales (aunque eso ya lo sabía, pues la nulidad estricta estaba descartada de antemano por tratarse de procedimientos diferentes). Sin embargo, también concluye que dichas tasas son esencialmente iguales, ya que confía en la calidad de unas estimaciones realizadas usando muestras tan enormes, y la diferencia verdadera parece ascender a aproximadamente a un 2% a favor de **N**, cifra que él considera de importancia marginal.

Ahora bien, el empleo del tratamiento novedoso cuesta alrededor de 40.000 euros anuales, mientras que el tratamiento **C** supondría un gasto de sólo 120 euros. Por otra parte, está descrito que el procedimiento más novedoso produce algunas reacciones adversas, las cuales no se registran para el convencional. Naturalmente, en este escenario, Traveller opta por éste último. Mientras hacía su análisis, va consolidando la convicción de que el problema es estructural y concierne a las pruebas de significación como tales. Un examen de las interioridades matemáticas de la prueba Ji-Cuadrado provee de un nuevo motivo para tal convicción. Al Si considerar genéricamente una tabla como las de este último ejemplo, se tiene lo que refleja la Tabla 4.

Tabla 4. Respuesta dicotómica (éxito o fracaso) a dos condiciones genéricas

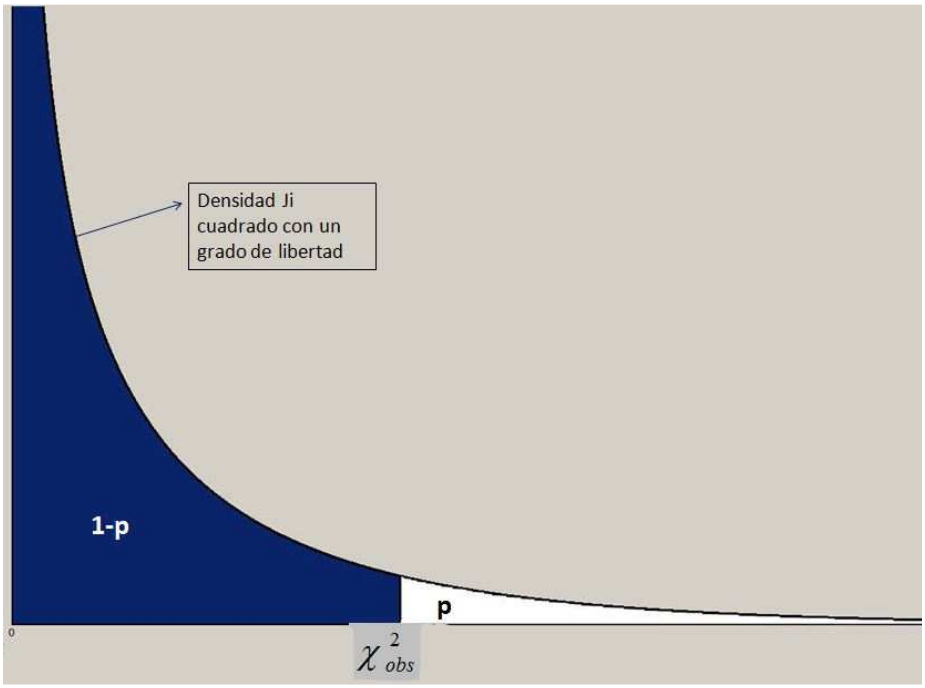
	Condición 1	Condición 2	**Total**
Fracaso	a	b	**a+b**
Éxito	c	d	**c+d**
Total	**n1**	**n2**	**n**

En este caso, p es igual al área a la derecha de χ^2_{obs} bajo la función de densidad de una variable que se distribuyen Ji-cuadrados con 1 grado de libertad, donde:

$$\chi^2_{obs} = \frac{n(ad-bc)^2}{n_1 n_2 (a+b)(c+d)}$$

Es fácil demostrar que siempre que las tasas observadas no sean exactamente iguales (en cuyo caso $\chi^2_{obs} = 0$), este valor crecerá en la medida que sea mayor el tamaño muestral n, se figura en el numerador. Siendo así (véase la Figura debajo), el valor de p decrecerá con dicho aumento.

Nuevamente, la aritmética del asunto confirma que, tomando muestras suficientemente grandes, siempre se hallará significación. Dicho sea de paso, esta es una gran noticia para la empresa transnacional que produce el fármaco en que se basa el tratamiento, la cual se puede permitir un estudio de grandes dimensiones.



Moraleja

Las pruebas de significación no aportan nada que no pueda derivarse de los intervalos de confianza; sin embargo, por lo general, simplemente, no aportan nada.

Nota:

El presente texto se ha inspirado en uno de los capítulos de un libro escrito por el estadístico británico radicado en Estados Unidos Andrew Vickers (*What is a P value anyway?*, 2010). Allí se esboza un problema similar al de los trayectos, pero a mi juicio, Vickers no lo desarrolla más que como una curiosidad y termina convalidando las pruebas de significación, aunque no se entiende por qué. Su conclusión es que hay que emplear dichas pruebas, pero con cautela.