Uso y abuso de pruebas de hipótesis

Federico Menéndez-Conde Lara*

March 16, 2014

Resumen

Se presenta un panorama del uso de pruebas de hipótesis en estudios clínicos, observando tanto algunas de sus virtudes como debilidades. En particular, se cuestiona la necesidad de realizar pruebas demasiado grandes.

Abstract

An overview of the use of hypothesis testing in clinical trials is presented, noting both some of its merits as some of its shortcomings. In particular, it is questioned the need to perform large trials.

Palabras clave: Pruebas clínicas, pruebas de hipótesis, medicina basada en evidencias, estadística inferencial aplicada

Keywords: Clinical trials, hypothesis testing, evidence based medicine, applied inferential statistics

---

Las pruebas aleatorias controladas son consideradas una de las herramientas más poderosas en la medicina contemporánea. En la literatura médica suelen ser referidas como el estandard dorado de la medicina basada en evidencias. La base teórica de estas pruebas son objetos matemáticos abstractos conocidos como pruebas de hipótesis, que fueron desarrollados por Ronald Fisher y Karl Pearson, entre otros notables matemáticos, a principios del siglo XX. En este artículo, ofrecemos una visión general de estas pruebas, y de como son aplicadas en los estudios clínicos. Comentamos tanto sobre las virtudes y utilidad de estas pruebas, como de algunos problemas inherentes, tanto teóricos como prácticos, de las mismas.

*Centro de Investigación en Matemáticas, Universidad Autónoma del Estado de Hidalgo

1 Qué son y cómo funcionan las pruebas de hipótesis

Las pruebas de hipótesis son poderosas herramientas de estadística inferencial, que son aplicadas en una gran variedad de áreas del conocimiento. En términos generales, una prueba de hipótesis consiste en establecer un criterio para decidir si se debe rechazar o no cierta hipótesis acerca de la población estudiada (la llamada *hipótesis nula*) en favor de otra (la *hipótesis alternativa*). Por lo general, en la práctica la hipótesis alternativa es la hipótesis de trabajo que los investigadores buscan probar como verdadera. La decisión que se adopte depende de los valores observados en una muestra seleccionada al azar (una *muestra aleatoria*); establecer el criterio de decisión, consiste en determinar una *región de confianza*, de manera que se rechazará la hipótesis nula en el caso de que los valores observados caigan dentro de dicha región.

Un aspecto fundamental a tomar en cuenta es que, al tomar decisiones a partir del conocimiento de datos de solo una pequeña parte de la población, es imposible evitar por completo la posibilidad de adoptar la decisión equivocada; ya sea porque se haya rechazado la hipótesis nula siendo que es verdadera, o porque no se rechace siendo que es falsa. Cuando ocurre la primera de estas situaciones, se dice que se ha cometido un *error de tipo I*, y si ocurre la segunda se dice que hubo un *error de tipo II*. Un objetivo natural es el minimizar la probabilidad de cometer estos errores. Sin embargo, como puede uno imaginarse, el disminuir el riesgo de cometer un error del tipo I aumenta el riesgo de cometer un error del tipo II, y viceversa; esto entraña una dificultad inherente. En la práctica, sucede que en la mayoría de las situaciones (como en las pruebas clínicas que discutiremos en la siguiente sección) se prioriza el evitar cometer un error del tipo I. A la probabilidad de cometer este tipo de error se le conoce como el *valor de significancia de la prueba*, y se suele denotar por la letra p. Mientras más pequeño sea el valor de p, se dice que la prueba de hipótesis (o "la evidencia") es más *significativa*. Si resulta ser que el valor de p es menor que un valor de significancia aceptable α (que, al menos en teoría, debería fijarse antes de realizar la prueba) se considera que la prueba es suficientemente significativa y se rechaza la hipótesis nula. En otras palabras, el valor de α determina la región de rechazo, y el que $p \leq \alpha$ es equivalente a que los datos obtenidos hayan caído dentro de dicha región.

Sin duda, todo esto pudiera sonar un tanto confuso para quien no esté familia–rizado con el tema; creemos que lo más apropiado para aclararlo, es usar un ejemplo sencillo, como el que presentamos a continuación.

Ejemplo uno. Imaginemos un agricultor que cosecha sandías con un peso promedio de 7 kilos. Le han ofrecido un fertilizante que, supuestamente, aumentará de manera considerable el tamaño de sus frutas. El agricultor decide poner

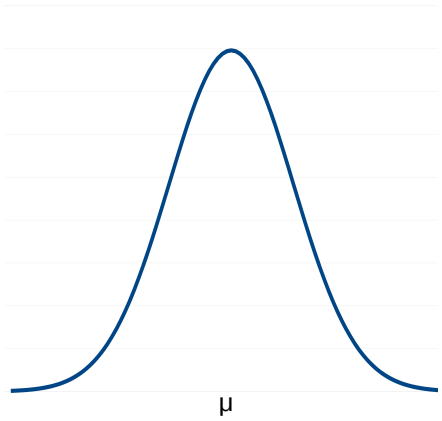

FIGURA 1. Gráfica de una distribución normal con media μ y σ=2.

a prueba el fertilizante por una temporada. Como el producto es costoso, decide que solo valdrá la pena usarlo a largo plazo si produce sandías de más de 8 kilos. Supongamos que este agricultor conoce de estadística, de manera que plantea una prueba de hipótesis para abordar su problema. Denotando por μ al peso promedio de las sandías producidas con el fertilizante, establece las siguientes hipótesis nula y alternativa:

$$\begin{aligned} \text{Hipótesis nula} \quad & H_0 : \ \mu \leq 8 \\ \text{Hipótesis alternativa} \quad & H_1 : \ \mu > 8 \end{aligned}$$

Desde luego, el agricultor quiere tener cierta garantía de que las sandías de verdad promediarán más de 8 kilos, por lo que propone un valor de significancia para la prueba de $\alpha = 0.15$; esto es, quiere estar al menos 85% seguro. Se asume también que el peso de las sandías está distribuido normalmente con media μ desconocida y desviación estándard $\sigma = 2$.[1] Esto significa que los pesos de las sandías se distribuyen según la gráfica de la figura 1. De una muestra de 50 sandías, el agricultor calcula el peso promedio $\bar{X}$ ($\bar{X}$ es lo que se llama un

[1]Cabe mencionar que esta es una fuerte suposición, que aquí tomamos por simplicidad y con fines ilustrativos.

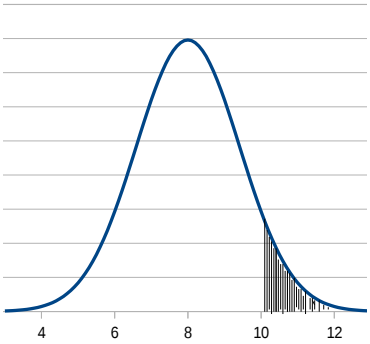

Figura 2. Distribución normal con
μ=8 y σ=2. El área sombreada es
el 15% del área total bajo la curva.
El intervalo correspondiente es
la región de rechazo para $p=0.15$.

estimador para μ). Si el valor de $\bar{X}$ está dentro del *intervalo de confianza*,[2] se rechaza la hipótesis H_0 en favor de H_1 y se compra el fertilizante. Ahora bien, dicho intervalo de confianza está determinado por valor de significancia aceptable propuesto. De forma más precisa: partiendo de suponer $\mu = 8$ (el máximo de los pesos promedios "no aceptables") el intervalo de confianza comienza a partir del valor de $\bar{X}$ para el cual

$$P(\bar{X} \geq x_0) \leq 0.15$$

(es decir, que la probabilidad de que el promedio sea mayor que x_0 es menor o igual que 15%). De manera gráfica, esto significa que el intervalo de rechazo comienza a partir del punto x_0 a partir del cual el área bajo la gráfica corresponde al 15% del área total bajo la curva (ver figura 2). En este caso, $x_0 = 10.08$ y por lo tanto el agricultor adquirirá el fertilizante si el promedio de las sandías que pesó es de al menos de 10.08 kilos.

Si H_0 fuera cierta, entonces la probabilidad de que $\bar{X}$ pese más de 10.08 kilos es pequeña (nótese que lo que "pequeña" quiere decir, depende del valor de α).

Remarcamos algunos aspectos pertinentes de este ejemplo, y de las pruebas de hipótesis en general:

[2] Al usarse un estimador con valores reales, la región de confianza es un intervalo; esta es a la vez la situación más sencilla y la más usual.

- En el ejemplo anterior se parte de suponer que los pesos de las sandías se distribuyen como en la figura 1. En principio, no sabemos que la realidad sea así y es solo una de muchas posibilidades. El criterio de decisión es dependiente de esta elección. Existen consideraciones teóricas que hacen que, en ciertas situaciones, la elección de la distribución normal sea una opción razonable (e.g. Hogg, McKean, Craig 2005).

- Más arbitraria todavía es la elección de la desviación estándard $\sigma = 2$, que solo fijamos con fines ilustrativos. De manera intuitiva, σ es un parámetro que indica la variabilidad de la población: mientras más grande sea σ, los valores se alejan con más frecuencia del promedio. En la distribución normal, la figura de la campana se hace gorda y chaparra cuando σ es grande, y se hace alta y puntiaguada si σ es pequeña.

- Además de la distribución normal, otras distribuciones usadas con frecuencia en medicina son la distrubución chi–cuadrada, la distribución t de student, y las distribuciones z.

- Una cuestión en extremo complicada en la práctica es el garantizar que, en la medida de lo posible, las selección de la muestra sea en realidad resultado de un procedimiento aleatorio. Si la elección no depende de forma exclusiva del azar, de modo que todo individuo de la población tenga la misma probabilidad de ser seleccionado, las consideraciones teóricas de la prueba de hipótesis no son válidas.

- El tamaño del valor de significancia p no cuantifica el valor real del parámetro. El tener un valor de p muy pequeño solo quiere decir que hay una muy alta probabilidad de que la hipótesis nula sea falsa; pero pudiera ser que fuera falsa "por muy poco".

El último de los puntos mencionados arriba, con mucha frecuencia es pasado por alto. Un valor de significancia alto es a veces entendido como evidencia de un efecto grande o importante; en realidad, el más modesto de los efectos puede ser detectado con una significancia tan grande como se quiera. Para conseguir esto, basta considerar muestras de tamaño grande. El entender y tener presente esto, nos parece fundamental para una adecuada comprensión e interpretación de las pruebas de hipótesis. El siguiente ejemplo y su discución posterior tiene como propósito explicar este punto.

Ejemplo dos. En cierta bodega muy grande se guardan, revueltas, muchas canicas de dos colores (digamos mitad verdes y mitad rojas). Aparte del color, las canicas son idénticas, en forma, peso y tamaño. Supongamos además que, por alguna razón, las canicas verdes valen más que las rojas; sabiendo esto, el desho-

nesto encargado de la bodega se ha dado a la tarea de sustraer cada día una pequeña cantidad de canicas verdes, reemplazándolas por canicas rojas. Después de cierto tiempo, hay 51% de canicas rojas y 49% de canicas verdes. Una supervisora entra en sospechas, y como no puede contar todas las canicas (hay millones de ellas), y el pesarlas no le daría información (pesan lo mismo), decide aplicar una prueba de hipótesis. Digamos que plantea que un valor de suficiencia de $\alpha = 0.05$ le basta (o sea, quiere estar al menos 95% segura). La hipótesis nula es: "hay la misma cantidad de canicas verdes que canicas rojas", siendo la alternativa: "hay más canicas rojas que verdes". Por la información que tenemos, nosotros sabemos que debería rechazar la hipótesis nula: aunque sea por poquito, hay más canicas rojas que verdes. No hacerlo implicaría que se está cometiendo un error del tipo II. La pregunta es entonces: ¿Cuántas canicas necesita sacar para rechazar la hipótesis nula? Es claro que la respuesta a esta pregunta no puede darse en términos absolutos, sino probabilísticos; para cada tamaño de muestra hay la posibilidad de que la hipótesis se rechace o de que no. Las probabilidades de que la supervisora capture al encargado (con su 95% de significancia) se ilustran en las dos primeras columnas de la tabla 1.[3] En las otras columnas, mostramos que para mayores niveles de significancia también es posible (de hecho casi seguro) detectar el cambio, bastanto tomar muestras algo más grandes . El cambio será detectado con una probabilidad de más de diez mil a uno con una muestra de 350 pelotas, incluso para niveles de significancia extremadamente altos de $p < 0.001$.

- Es muy importante notar que, en todas estas tablas, la cantidad total de canicas rojas y verdes no cambia. Aunque los valores de significancia sí lo hacen.
- Sin cambios en la población general, para obtener un valor de significancia más pequeño (tan pequeño como se quiera, en realidad) basta tomar muestras más grandes. Es en este sentido que las pruebas de hipótesis son herramientas matemáticas muy efectivas: permiten detectar cambios muy pequeños de manera significativa, siempre y cuando la muestra tomada sea suficientemente grande.

2 Pruebas clínicas

En medicina, las pruebas de hipótesis son una herramienta usada con mucha frecuencia. Principalmente, su aplicación consiste en evaluar nuevos tratamientos, comparándolos con otros previamente existentes, o con algún placebo. En este

[3] Se usaron aproximaciones normales con $p = 0.5$ y $\sigma^2 = 0.25/n$ para calcular la región de rechazo; y con $p = 0.51$, $\sigma^2 = (0.51 * 0.49)/n$ para calcular la probabilidad de caer ahí.

Tamaño de la muestra	Probabilidad de rechazar H_0 para $\alpha = 0.05$	Tamaño de la muestra	Probabilidad de rechazar H_0 para $\alpha = 0.01$	Tamaño de la muestra	Probabilidad de rechazar H_0 para $\alpha = 0.001$
50	25.8%	100	36.7%	150	45%
100	63.8%	150	74.5%	200	81%
150	91.2%	200	95%	250	97.1%
200	99.1%	250	99.6%	300	99.8%
250	$> 99.9\%$	300	$> 99.9\%$	350	$> 99.99\%$

Table 1: Probabilidades de rechazo de la hipótesis nula H_0 en la situación del ejemplo dos.

contexto, las pruebas de hipótesis se conocen con el nombre de *pruebas aleatorias controladas*.[4] Estas pruebas se consideran la piedra angular de la llamada *medicina basada en evidencias*. En términos generales consisten en elegir de manera aleatoria un conjunto de pacientes, y dividirlos – también de manera aleatoria, desde luego – en dos o más grupos. A cada grupo de pacientes se asigna un tratamiento distinto, y se comparan los resultados de cada uno de ellos. Por ejemplo, una situación usual es con solo dos grupos: al primero se le asigna el tratamiento a probar (digamos, una nueva medicina de la que se quiere mostrar su eficacia) mientras que al otro se le proporciona un placebo. En dicho caso, la hipótesis nula será que el nuevo medicamento y el placebo funcionan igual, mientras que la hipótesis alternativa será que el medicamento da un mejor resultado que el placebo. Es claro que un requisito indispensable para que estas pruebas sean válidas, además de la aleatoriedad de la muestra, es que el procedimiento se haga a ciegas: es decir, un paciente no debe saber a cuál de los grupos pertenece (ni tener elementos de sospecha en uno u otro sentido); esto no es nada sencillo de conseguir en la práctica, en especial si los médicos participantes sí conocen a qué grupo está aignado cada paciente. Debido a esto último, idealmente se espera que los médicos tampoco tengan conocimiento sobre si están administrando el medicamento o el placebo; en este caso, se dice que la prueba es a *doble ciego*. Las pruebas aleatorias controladas a doble ciego constituyen el estándard dorado en pruebas clínicas para nuevas medicinas.

[4]El nombre en inglés *randomized controlled trials* es mucho más común.

2.1 Un poco de historia

Es en general aceptado que la primera prueba aleatoria controlada, de acuerdo a los parámetros modernos, fue una prueba realizada en 1947 en siete hospitales de Inglaterra para evaluar la efectividad de la estreptomicina en el tratamiento de la tuberculosis pulmonar. Los resultados fueron publicados en 1948 en el British Medical Journal (Medical Research Council, 1948). De una población de 107 pacientes, a un grupo de 55 se les administró un tratamiento con estreptomicina durante cuatro meses; los restantes 52 formaron el grupo de control. El proceso de selección se hizo mediante la generación de números aleatorios, y la asignación de los grupos se mantuvo oculta tanto para los médicos como para los pacientes durante todo el proceso; es decir, se trató de un estudio a doble ciego. El periodo de observación fue de seis meses. Al final de ese periodo, habían fallecido cuatro pacientes del grupo de estreptomicina, y catorce del grupo de control (el 7% y el 27% respectivamente). Del grupo de estreptomicina, 28 pacientes presentaron una mejoría considerable (el 54%), mientras que ese fue el caso solo para cuatro de los pacientes del grupo de control (el 8%). Estos resultados son tan contundentes, que habría que ser demasiado escéptico para dudar de la eficacia de la estreptomicina después del primer vistazo. No parece necesario realizar minuciosos análisis estadísticos para convencernos de ello. Solo por dejar constancia, con esos datos se muestra que la estreptomicina evitó muertes para un nivel de significancia de $p = 0.02$, y que mejoró la condición de los pacientes para $p < 0.001$.[5] Estos niveles de significancia son muy altos, sobre todo si tomamos en cuenta lo pequeño de la muestra. La reducción de riesgo absoluto (RRA) de muerte en el periodo fue de 20%, mientras que la reducción de riesgo relativo *RRT* fue del 74%. Resulta importante observar que, estos no son los valores reales de reducción de riesgo en la población general, sino solo los valores que arrojan los estimadores puntuales en la muestra considerada.

En realidad, antes del estudio de estreptomicina descrito arriba ya se habían realizado numerosas pruebas clínicas que pudieran considerarse pruebas aleatorias controladas; sin embargo, la selección de los grupos de estudio se hacía mediante procedimientos que no pueden considerarse aleatorios en el sentido estricto. Estas pruebas "pseudoaleatorias" fueron muy frecuentes desde los años 1920's, influenciadas por el fuerte desarrollo de la teoría y práctica estadística que tuvo lugar a principios del siglo XX. En ellas, los procedimientos de selección de grupos incluían el ir asignando alternadamente a los pacientes, o el hacerlo lanzando un par de dados o una moneda. Un ejemplo de esto fue realizado por el mismo organismo británico[6] que realizó la prueba de estreptomicina: pocos años antes

[5] Usando prueba de distribución normal para diferencia de medias.

[6] MRC: Medical Research Council.

habían puesto a prueba un medicamento (patulin) para el tratamiento del resfriado común. La asignación de pacientes en esta prueba fue hecha alternando a los pacientes mediante un complicado método de asignación que involucraba el uso de unas tarjetas rotuladas. Cabe señalar que la prueba del patulin sí fue hecha a doble ciego, y que ya desde ese entonces resultaba muy clara la importancia de que las pruebas clínicas tuvieran dicha característica:

> "La experiencia previa nos ha convencido de que, en una prueba de esta naturaleza, es de gran importancia que se prevenga que tanto el personal médico como los pacientes puedan adivinar cuál de los dos tratamientos es genuino y cuál espurio". (Medical Research Council, 1944).

Los estudios clínicos controlados, si quitamos el ingrediente de la selección aleatoria y la aplicación de la moderna teoría estadística, se remontan varios siglos atrás. Hay referencias a ellos en diversas culturas de la antigüedad (un caso curioso se incluye en el Antiguo Testamento, en el Libro de David). Un ejemplo muy famoso, y de suprema importancia en la historia de la medicina, fue el experimento con el que a mediados del siglo XVIII, el médico inglés James Lind demostró que el consumo de cítricos curaba el escorbuto (Lind, 1753). A seis marineros, de un grupo de doce que padecían la enfermedad, dio una ración diaria de dos naranjas y un limón; a la otra mitad les administró uno de los supuestos tratamientos convencionales de la época (que incluía vinagre, sidra y agua de mar). Después de un tiempo, todos los marineros del grupo de los cítricos se había curado, mientras que todos los del otro grupo seguían enfermos. James Lind no reporta cómo decidió a qué marineros darles limón y a cuales no. El ejemplo más antiguo del que tenemos noticia en el que sí se describe de forma explícita el procedimiento (pseudoaleatorio) para ubicar pacientes en grupos de control, es un estudio hecho por Lohner (1835) en Nürenberg (cf. Stolberg, 2006), en el que se puso a prueba cierto medicamento homeopático. No resultará demasiado sorprendente el que el medicamento no haya superado la prueba. Este estudio tiene notable importancia histórica: además de lo ya señalado, este fue el primer experimento realizado a doble ciego del que se tiene registro; sin embargo, adoleció de un defecto: los participantes del experimento estaban informados de estar participando en el mismo. Esta situación puede producir sesgos, y no es lo ideal desde el punto de vista teórico. Es sin embargo, un mal necesario que comparten las modernas pruebas aleatorias controladas con el experimento de Nürenberg.

En el presente, las pruebas clínicas requieren del consentimiento de los pacientes participantes. Esto quedó establecido en la Declaración de Helsinki (World Medical Association, 1964), que establece principios éticos para estudios clínicos que involucran seres humanos. Desde mediados de la década de los 1970's, en

la práctica totalidad de las pruebas aleatorias controladas que se realizan, los pacientes considerados en el estudio son informados de los detalles del mismo, siendo su consentimiento explícito y por escrito un requisito indispensable para su participación. Si bien esto es, desde luego, lo correcto desde el punto de vista de la ética y el respeto a los pacientes, tiene al menos dos inconvenientes. Por un lado, no hay ninguna razón para pensar que la población formada por pacientes que dieron su consentimiento es homogénea con la población general (lo que implica un sesgo de no aleatoriedad en el muestreo). Por otro lado, el hecho de conocer del experimento, puede también producir sesgos en los resultados; por ejemplo, el que los pacientes estén informados de que tal vez están recibiendo un placebo, puediera hacer menos efectivo el "efecto placebo". Cabe hacer notar que en el estudio pionero de la estreptomicina de 1947, así como en muchos de los estudios realizados antes de la Declaración de Helsinki, no se tuvo ese problema; en esos casos, los pacientes (y a veces tampoco los médicos) no tenían conocimiento de estar participando en una prueba clínica con grupos de control. Puede decirse que, en ese sentido, estas pruebas fueron un tanto "más puras" que las realizadas hoy.

Desde el experimento de la estreptomicina de 1947, el uso de las pruebas aleatorias controladas se fue haciendo más y más frecuente. El gran número de pruebas que se realizan actualmente, hace demasiado complicado el poder determinar su número exacto; de acuerdo a algunas estimaciones el número ha llegado a superar las 25 mil pruebas anuales (Heneghan, 2010). El tamaño de las pruebas también ha ido creciendo enormemente, y es muy común que involucraren a miles de pacientes, siendo realizadas simultáneamente en varios centros de salud, incluso de países distintos. Desde los 1990's no es difícil encontrar ejemplos de pruebas que incluyen a decenas de miles de pacientes que son tratados durante años. La realización de estos estudios masivos resulta en extremo onerosa, llegando a costar cientos – y en casos hasta miles – de millones de dólares (Roy, 2010). Debido a esto, las pruebas clínicas de gran tamaño son en su gran mayoría financiadas y realizadas por compañías farmacéuticas, quienes cuentan con la perspectiva de recuperar la inversión al colocar nuevos medicamentos en el mercado. Algunas pruebas grandes son todavía financiadas con dinero público, pero su número se está reduciendo(Yusuf, 2004). Entre las pruebas de menor tamaño, de aproximadamente 150 a 3000 sujetos de estudio, es todavía relativamente común que sean realizados por grupos de investigación independientes.

2.2 El estudio HERS

La principal fortaleza de las pruebas de hipótesis, es que nos proveen de certeza (salvo niveles de significancia estadística) en situaciones en los que las obser-

vaciones empíricas y epidemiológicas dan resultados no convincentes o hasta aparentemente contradictiorios. Ocurre a veces que los alcances de las pruebas de hipótesis llegan aún más lejos: Se han dado casos en que pruebas de hipótesis han echado por tierra supuestos que se tenían por ciertos a partir de "evidencia epidemiológica contundente". Un maravilloso ejemplo de esto fue el ofrecido por el estudio *HERS (Heart and Estrogen/progestin Replacement Study)* (Hulley et al, 1998), una prueba aleatoria controlada a doble–ciego para verificar la eficacia del uso de hormonas para prevenir enfermedades cardiacas en mujeres post-menopáusicas. El diseño estadístico de la prueba, así como los detalles del muestreo fueron publicados en Grady et al (1998).

Durante toda la década de los 1990's, a millones de mujeres les fue recetado un *tratamiento de reemplazo hormonal* (HRT[7], por sus siglas en inglés), que consistía en unas tabletas con una combinación de estrógeno y progestina. Se tenía la creencia de que con ese tratamiento, se ayudaba a evitar en las mujeres menopáusicas los más diversos males asociados con la edad avanzada (senilidad, osteoporosis, diversos tipos de cáncer, etc.) Fue sobre todo muy extendida la creencia de que el HRT ayudaba de manera notable a reducir el riesgo de sufrir infartos. Esta creencia fue respaldada por numerosos estudios observacionales que parecían confirmar la teoría; en Hulley et al (1998) están citados doce de estos estudios. Habiendo mucha evidencia en favor de la utilidad del HRT para prevenir infartos, la compañía farmacéutica Wyeth decidió dejar asentado el tema de manera oficial, de manera que financió y condujo una prueba aleatoria controlada para comprobarlo. Esta prueba fue HERS.

Las pacientes que participaron en el estudio HERS fueron 2763 mujeres post–menopáusicas de entre 55 y 80 años, diagnosticadas con alguna enfermedad coronaria. A un grupo de 1380 de ellas se les administró HRT por poco más de cuatro años, mientras que al grupo de control de 1383 pacientes se les dio un placebo. Al final del estudio, fallecieron por enfermedad cardiaca 71 mujeres del grupo de HRT (el 5.1%), y solo 58 de las del grupo de control (el 4.2%); las muertes totales fueron 131 en el grupo de HRT y 123 en el grupo de control (9.5% y 8.9% respectivamente). De esta forma, tanto las muertes por enfermedades cardiacas como las muertes totales aumentaron con el tratamiento de reemplazo hormonal; hay que decir que las diferencias fueron muy pequeñas para ser significativas).[8] También se midieron la incidencia de diversos tipos de cáncer y de fracturas por osteoporosis; en ningún caso los resultados mostraron cambios significativos en uno u otro sentido. Todo esto contradecía lo esperado por Wyeth, y lo que habían mostrado los estudios epidemiológicos. Había que desechar la teoría. El proceso

[7]Hormone Replacement Therapy

[8]En medicina es usual considerar como significativo un valor de p menor que 0.05.

de desecharla resultó sin embargo lento en extremo. Resulta curioso (o el calificativo que se prefiera) el hecho de que ya bien entrado el siglo XXI, ignorando los resultados de su propia prueba, Wyeth seguía promocionando el uso de HRT para la prevención de enfermedades cardiacas; todavía más grave es el hecho de que en el año 2000, dos años después de publicados los resultados del HERS, el *American College of Obstetricians and Gynecologists* seguía recomendando el uso del HRT para prevenir infartos. Detalles de esta historia pueden consultarse en Monihan y Cassels (2005).

El estudio HERS no fue suficiente para reducir de manera importante las prescripciones de HRT. El golpe de gracia llegó en 2002, con la publicación de los resultados de otra prueba aleatoria controlada: WHI (Women's Health Initiative). Esta fue una prueba aleatoria controlada masiva, financiada con dinéro público (del gobierno de Estados Unidos) en la que participaron 16608 mujeres postmenopáusicas sanas durante más de cinco años. En este periodo, aumentaron en el grupo de HRT los infartos al miocardio, los accidentes cerebrovasculares y los casos de cáncer de mama. Se redujeron los casos de cáncer de colon y las fracturas por osteoporosis. Es muy importante aclarar que los cambios en la incidencia de cada uno de estos males – tanto los que aumentaron como los que se redujeron – son marginales, muy pequeños en cuanto a reducción de riesgo absoluto (aunque estadísticamente significativos para $p = 0.05$).

En la actualidad, el tratamiento de reemplazo hormonal está indicado solamente para aliviar algunos de los síntomas de la menopausia, y no suele administrarse a largo plazo. En muchos países está contraindicado para la población considerada con alto riesgo de padecer enfermedades cardivasculares. Para aliviar síntomas de la menopausia, la eficacia clínica de la terapia de reemplazo hormonal está más que demostrada. Además, a diferencia de los silenciosos riesgos a largo plazo que requieren de pruebas de hipótesis para probarse, sus beneficios pueden ser comprobados de manera inmediata por las pacientes.

2.3 Una cuestón paradójica

Como hemos mencionado, desde hace ya algún tiempo, las pruebas clínicas que se realizan, llegan a ser de gran tamaño; en ocasiones, incluyen a decenas de miles de participantes, pudiendo encontrarse estudios con más de cien mil pacientes, como por ejemplo el de Ulmer et al (2004). En años recientes, los costos de realizar estas pruebas se han elevado en gran medida, no solo por el tamaño y la duración de las mismas, si no también porque las regulaciones para realizarlas han aumentado, lo que de acuerdo a Yusuf (2004) y Califf (2006) entre otros autores, ha redundado en complicaciones administrativas y provocado un sinnúmero de costos añadidos. Se

ha señalado la necesidad de seguir realizando pruebas grandes y multicéntricas, de preferencia financiadas por gobiernos – y no por la industria farmacéutica – por lo que resulta necesario reducir costos. La siguiente cita, del Dr. Salim Yusuf, médico canadiense que ha dirigido grandes pruebas clínicas multicéntricas da un argumento sobre la importancia de continuar realizando dichas pruebas (Yusuf, 2004):

> "... para la mayoría de las enfermedades crónicas (e.g., enfermedad cardiovascular o cáncer) los tratamientos suelen tener cuando mucho un efecto benéfico moderado (...) Esta situación transformó las pruebas aleatorias controladas, conduciendo a la creación de pruebas grandes, multicéntricas, con frecuencia involucrando a decenas de miles de sujetos"

De otro lado del debate, existen objeciones de lo más diversas que cuestionan el que las pruebas aleatorias randomizadas sean en verdad apropiadas y útiles en muchas de las circunstancias en que son aplicadas; en esta línea están trabajos como el de Hunink (2004), el de Grossman y McKenzie (2005) y el de Ventegodt et al (2009). Acá queremos considerar un punto que no cuestiona la utilidad de las pruebas de hipótesis en sí, pero sí la necesidad de implementar pruebas grandes. Se trata de una cuestión muy simple y directa, pero un tanto paradójica:

> *Mientras menos eficiente es un tratamiento, se requiere de pruebas aleatorias más grandes para detectar su utilidad.*

Esto es lo señalado en la cita de S. Yusuf, aunque visto desde una perspectiva muy distinta. En otras palabras, las pruebas que involucran a muchos miles de pacientes son usadas (invariablemente) para evaluar tratamientos con un "efecto benéfico moderado". Como en nuestro ejemplo de las pelotas verdes y rojas, una mínima diferencia podrá ser fácilmente detectada con pruebas de hipótesis suficientemente grandes. Cuando un medicamento es muy eficaz, no es necesario usar muestras grandes para probarlo. En el ejemplo pionero de la estreptomicina, poco más de cien pacientes bastaron para arrojar resultados significativos. Difícilmente se requeriría de más de una veintena de pacientes para probar la eficacia de la penicilina (con un nivel de significancia $p < 0.001$). A James Lind le bastó observar a doce pacientes para sacar conclusiones; pero es que ahora sabemos que la vitamina C es eficaz siempre contra el escorbuto. Si un medicamento es por completo eficaz, después de diez éxitos seguidos tendríamos ya una garantía con $p < 0.001$ de que su uso supera al placebo. De esta manera, el uso de pruebas de hipótesis masivas, significa casi siempre que el tratamiento que se está probando, aspira a llegar a beneficiar solo a una pequeña fracción de los pacientes a los que

les será recetado. En vista de esto, tal vez resulte sorprendente el que se sigan llevando a cabo muchas pruebas masivas de medicamentos, y el que se inviertan muchos millones de dólares en las mismas. El por qué se siguen realizando estas pruebas es muy simple: los medicamentos que resultan aprobados a partir de estas pruebas por lo general resultan en un muy productivo negocio para las compañías farmacéuticas, quienes suelen recuperar con creces la inversión; realizar las pruebas es entonces, podríamos decir, el comportamiento lógico de una industria saludable con fines de lucro. Ahora bien, podemos preguntarnos: ¿cómo es que medicamentos que solo ofrecen un beneficio moderado resultan tan buen negocio? La respuesta es otra vez simple, y otra vez aparentemente contradictoria:

> *Mientras menos eficiente sea un tratamiento, hay que dárselo a más personas, para curar al mismo número dado de pacientes*.

En otras palabras: mientras menos eficiente sea un medicamento hay que venderlo a más personas. Sobre todo a personas sanas que en realidad no requieren ni del medicamento que se les administra ni de ningún otro. Así de extraño (o maquiavélico) como pueda sonar eso, muchos de los medicamentos más usados en la actualidad (y los que reditúan más beneficios a las compañías farmacéuticas) son vendidos a millones de personas en periodos de tiempo extensos, de las cuales solo a algunos pocas (proporcionalmente hablando) les reportará algún beneficio. Para ilustrar este hecho, presentamos la siguiente cita del Prof. Rory Collins, médico y epidemiólogo británico:

> "Si ahora, a un añadido de 10 millones de personas en el mundo, con alto riesgo [de morir de infarto] se le pone en tratamiento con estatinas, esto salvaría alrededor de 50 mil vidas por año [debidas a causas cardiovasculares]"

Esta cita está tomada de un comunicado de prensa de la British Heart Foundation (Collins, 2004), acerca de un reporte del Heart Protection Study, una prueba aleatoria controlada (conducida por el mismo R. Collins, entre otros) en la que se evaluó la eficacia de la simvastatina en 20536 pacientes con alto riesgo de padecer enfermedades cardiovasculares (Heart Protection Study Collaborative Group, 2002). Salvar 50 mil vidas (o de forma más correcta: prolongar 50 mil vidas) suena muy bien, y hasta espectacular. Pero es el 0.5% de la población de diez millones hipotéticamente tratada. Uno de cada doscientos. Y si nos ponemos un poco inquisitivos surgen preguntas bastante pertinentes: ¿Cuánto tiempo se prolonga la vida de ese 0.5%? ¿En verdad se prolonga la vida, o solo se reducen las muertes por causas cardivasculares? ¿Cómo se compara ese resultado con el uso de otros tratamientos? Etcétera. Consideramos que para los pacientes (los

hipotéticos diez millones) es fundamental tener toda la información relevante al medicamento que le es administrado. Y si se sabe que la probabilidad de que llevar un tratamiento dado funciona en el 0.5% (o en el 10%, o en el 50% o lo que sea) el paciente debería de ser informado. Ya será cuestión del libre albedrío, la idiosincrasia, el carácter, el criterio y la reacción emocional de cada paciente si decide tomar el tratamiento o no. En esta línea está la siguiente opinión:

> "Hombres en alto riesgo [de infarto] en edades de 30 a 69 años de edad, deben ser advertidos que alrededor de 50 pacientes deben ser tratados [con estatinas] por 5 años para prevenir un evento [infarto o muerte cardiovascular]. En nuestra experiencia, muchos hombres eligen no tomar la estatina ante esta evidencia..." (Abramson y Wright, 2007)

Este tipo de información muy rara vez es presentada a los pacientes.

En el estudio WHI que mencionamos en el apartado anterior, se mostró que el tratamiento de reemplazo hormonal HRT aumentaba el riesgo de presentar alguna condición coronaria grave (se consideraron infartos al miocardio y muertes cardiovasculares). Pero el aumento de los casos fue en realidad muy pequeño: un 1.9% de casos en el grupo de HRT, por un 1.5% en el grupo de control; el gran tamaño de la muestra implicó que este aumento de riesgo absoluto de 0.4% fuera estadísticamente significativo.[9] Se entiende que el HRT beneficia a muchas mujeres, al aliviar algunos síntomas de la menopausia, mejorando su calidad de vida en lo inmediato. ¿Es correcto vetarles de su uso? ¿No sería más apropiado dejar la decisión en cada paciente? Si a una paciente con alto riesgo de infarto se le informa que el llevar por periodos largos un tratamiento hormonal aumenta un poco (solo un poco) la probabilidad de que sufra un infarto dentro de algunos años, ¿tomará el tratamiento? Eso dependería desde luego ya de cada quién. Sin embargo, resulta que en muchos países el HRT está contraindicado para personas con factores de riesgo para enfermedades cardiovasculares. Porque una prueba aleatoria controlada masiva mostró un pequeño aumento en el riesgo. Es claro que el estudio HERS, el WHI y otros dejan muy claro que el HRT está muy lejos de ser la panacea de la salud que se pensaba hace algunos años; pero de ahí a prescribir un tratamiento que – en algunos casos y en situaciones puntuales – puede ser de utilidad, hay un largo trecho. También es pertinente señalar que, en el WHI, la mortalidad total en el grupo de control fue de 2.069%, mientras que en el de HRT fue de 2.716%;[10] este aumento del 0.026% no fue ni siquiera considerada estadísticamente significativo, a pesar del enorme tamaño de la muestra.

[9] Se reportaron 164 casos de entre las 8506 pacientes en el grupo del tratamiento, y 122 de 8102 en el grupo de control.

[10] 218 y 231 casos, respectivamente.

Preguntémonos ahora, ¿qué pasaría si los resultados del WHI sobre HRT y enfermedades coronarias hubieran salido al revés? Es decir, supongamos que en vez de mostrar que el HRT aumenta marginalmente el riesgo de infarto, se hubiera mostrado que lo reduce (también marginalmente). Sin temor a equivocarnos, lo que hubiera sucedido bajo ese supuesto, es que el HRT se recetaría de manera indiscriminada como una "importante medida preventiva" a millones de personas, la mayoría mujeres sanas sin riesgo real de infarto (y quizá hasta a muchas mujeres demasiado jóvenes para haber llegado a la menopausia). Eso es lo que suele ocurrir en esos casos. En el mundo de la interpretación de las pruebas de hipótesis no suele haber medias tintas: Hay ganadores y perdedores absolutos. Cuando un medicamento "pasa una prueba aleatoria controlada", por lo general no se da mucha importancia al hecho de que los beneficios encontrados hayan sido mínimos, o el que se haya requerido una prueba muy grande para detectarlos. Es común escuchar como argumento a favor de la eficacia algún medicamento el que haya pasado una prueba en que se evaluaron "muchos miles de pacientes"; está claro que ese debería ser más bien un argumento en contra de su eficacia. En resumidas cuentas, lo único que importa es que sean los resultados sean "significativos". Esta situación no es de ninguna manera un problema de las pruebas de hipótesis en sí (que, al menos en teoría y si su aplicación es la adecuada, son objetos matemáticos que dan información precisa de los hechos). Es más bien una cuestión de mercadotecnia. Uno de los ejemplos más notables de esto es el que nos ofrecen las estatinas, medicamentos recetados para la prevención de infartos al miocardio y enfermedades cardiovasculares en general. Se han realizado varias docenas de pruebas aleatorias controladas con estatinas, considerando diversos grupos de poblacionales. De todos esos estudios, el que ha mostrado la mayor reducción de riesgo absoluto de infarto fue una prueba conocida como *4S: Scandinavian Simvastatin Survival Study* (Scandinavian Simvastatin Study Group, 1994). Fue una prueba aleatoria controlada a doble ciego, financiada por la compañía farmacéutica Merck, en el que se comparó la eficiencia de la simvastatina contra un placebo. La población de estudio considerada fueron pacientes que sufrieran de angina de pecho o que hubieran tenido previamente un infarto al miocardio. Fueron considerados un total de 4444 pacientes (3617 hombres y 827 mujeres). El tiempo de observación fue de 5.4 años.[11] Al final del periodo de observación, habían fallecido el 9.3% de los pacientes en el grupo de control, y el 6.1% de los del grupo de la simvastatina; esta reducción de riesgo absoluto del 3.2% es la más grande jamás registrada en la historia de las prueba aleatorias controladas con estatinas.[12] Como apunte al margen, entre las mujeres que par-

[11] Esa es la mediana de los tiempos de observación, que abarcaron entre 4.9 y 6.3 años entre los sobrevivientes.

[12] Desde luego, la cifra de reducción de riesgo relativo de 34% resulta mucho más llamativa.

ticiparon en el 4*S*, hubo menos fallecimientos en el grupo de control que en el de simvastatina.

2.4 Sobre la presentación de resultados

Cuando se presentan los resultados de las pruebas clínicas, la mayoría de las veces se hace en términos de riesgos relativos, omitiendo mencionar el riesgo absoluto; la necesidad de usar este último, tanto en los reportes como en la comunicación médico–paciente, ha sido señalada en numerosas ocasiones; por ejemplo ver Edwards et al (2002), Menéndez-Conde Saiz (2003), Paling (2003), Abramson y Wright (2007), y Kaplan y Victor (2009).[13] Es claro que tanto la reducción de riesgo relativo RRR como la reducción de riesgo absoluto RRA (o equivalentemente a este último, el número necesario a tratar NNT=1/RRA) son índices válidos que pueden ofrecer información relevante; siempre y cuando se presenten en el contexto adecuado, y se deje en claro cuál es el tipo de reducción de riesgo del que se está hablando. Si leemos o escuchamos que un tratamiento "reduce el riesgo en un 20%" sin hacer esa aclaración son palabras huecas sin un significado real; apropiadas en propaganda comercial, pero no en una consulta médica. El que un paciente sepa que la RRR es de 20% no le será de mucha utilidad por sí solo, puesto que lo mismo puede tratarse de que el riesgo se reduce de un 80% a un 64%, que una reducción del 1.5% al 1.2%; situaciones dramáticamente distintas, claro está. En la primera situación el paciente tiene un 20% de probabilidad de ser beneficiado por el tratamiento, mientras que la segunda esto se reduce al 0.3%; estos números corresponden a la RRA. Al dar al paciente una valoración concreta de la probabilidad de ser beneficiado por el tratamiento, el índice RRA resulta ser más apropiado. Sin embargo, por sí solo tampoco da la información completa: no es lo mismo reducir un riesgo de 2% a 1%, que hacerlo de 80% a 79%; en ambos casos la RRA es de 1%. Sin duda, lo correcto es ofrecer la información completa de la manera más clara posible. La RRA, el NNT y la RRR son índices que usados correctamente pueden ayudar a este fin; usados de manera aislada o descontextualizada, tienen el efecto contrario de desinformar. Al margen de esto, un aspecto teórico a tomar en cuenta con estos índices es que no son valores absolutos: No podrían serlo, al ser dependientes de los resultados de una muestra. En forma estricta, tanto la RRA como la RRR tendrían que ser presentados como los puntos medios de intervalos de confianza específicos, dependientes de valores de significancia. Claro está que pretender hacer eso en la práctica clínica sería utópico, ilusorio, confuso e innecesario; presentar estos índices como valores puntuales es sin duda más que suficiente.

[13] Explicaciones de estos términos, dirigidas al público general, pueden consultarse, por ejemplo en Scott (2008) y en Harris y Taylor (2009).

A manera de conclusión, las pruebas de hipótesis son poderosas herramientas estadísticas que han reportado gran utilidad en la práctica médica, ofreciendo información valiosa. Sin embargo, una incompleta comprensión de las mismas (o peor aún, una alevosa presentación parcial de los datos), puede llevar a inadecuadas interpretaciones de los resultados.

Bibliografía

Abramson, J. y Wright, JM. (2007). Are lipid lowering guidelines evidence based? *Lancet*, Vol 369, pp 168-169.

Califf, R. (2006). Clinical trials bureaucracy: unintended consequences of well-intentioned policy. *Clinical Trials*, Vol. 3 no. 6, pp. 496-502.

Collins, R. (2004). LIFE-SAVER: World's largest cholesterol-lowering trial reveals massive benefits for high-risk patients. *MRC/BHF Press Release*.

Edwards, A., Elwyn, G. y Mulley, A. (2002). Explaining risks: turning numerical data into meaningful pictures. *British Medical Journal*, 324:827–30.

Grady D. et al. (1998). Heart and estrogen/progestin replacement study (HERS): design, methods and baseline characteristics. *Controlled Clinical Trials*, 19:314-335.

Grossman, J. y McKenzie, F. (2005). The randomized controlled trial: gold standard, or merely standard? *Perspectives in Biology and Medicine*, 48(4):516-34.

Harris, M. y Taylor, G. (2009). *Medical and Health Science Statistics made Easy*, 2a Edición, Jones & Bartlett Pub.

Heart Protection Study Collaborative Group (2002). MRC/BHF heart protection study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet*, 360 (9326):7-22.

Heneghan, C. (2010). How many randomized clinical trials are published each year? Recuperado de http://blogs.trusttheevidence.net/carl-heneghan/how-many-randomized-trials-are-published-each-year

Hogg, R., McKean, J. y Craig A. (2005) *An Introduction to Mathematical Statistics*, New Jersey: Pearson Prentice Hall, Sexta Edición.

Hulley, S. et al. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association*, Vol. 280, No. 7, pp. 605-13.

Hunink, M. (2004). Does evidence based medicine do more good than harm? *British Medical Journal*, 329:1051.1.

Kaplan, N. y Victor, R. (2009) *Clinical Hypertension*. Lippincott Williams & Wilkins, Décima Edición.

Lind, J. (1753). *A Treatise of the Scurvy*, London: A. Millar.

Löhner, G. (1835). Die homöpathischen Kochsalzversuche zu Nürnberg. *Allgemeine Zeitung von und für Bayern*, 1835.

Menéndez-Conde Saiz, M. (2003) Medicina basada en evidencias. Reducción relativa vs reducción absoluta de riesgo. *Revista Mexicana de Cardiología*, Vol. 14, Num. 2, pp 57 - 60.

Monihan, R. y Cassels, A. (2005) *Selling Sickness*, New York: Nation Books Avalon Pub. 2005.

Paling, J. (2003). Strategies to help patients understand risks. *British Medical Journal* 327:745.

MRC Patulin Clinical Trials Comittee (1944). Patulin in the tratment of common cold. *Lancet*, 2:373-5.

MRC Streptomycin in Tuberculosis Trials Comittee (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, ii:769–783.

Rossouw, JE. et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association*, 288(3):321-33.

Roy A. (2012). Stifling new cures: The true cost of lengthy clinical drug trials. *ProjectFDA Report*, No. 5.

Scandinavian Simvastatin Study Group (1994). Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study. *Lancet*, Vol. 344, pp 1383–1389.

Scott, I. (2008). Interpreting risks and ratios in therapy trials. *Australian Prescriber*, 31:12-6.

Stolberg, M. (2006). Inventing the randomized double–blind trial: the Nuremberg salt test of 1835. *Journal of the Royal Society of Medicine*, Vol. 99, pp 642–643.

Ulmer, H. et al. (2004). Why Eve is not Adam: Prospective follow-Up in 149,650 women and men of cholesterol and other risk factors related to cardiovascular and all-cause mortality. *Journal of Women's Health*, Vol. 13, No.1, pp. 41–53.

Ventegodt S. et al. (2009). Evidence-based medicine: four fundamental problems with the randomized clinical trial (RCT) used to document chemical medicine. *International Journal of Adolescent Medicine and Health*, 21(4):485-96.

World Medical Association (1964), *World Medical Association Declaration of Helsinki*. Recuperado de http://www.wma.net/en/30publications/10policies/b3/index.html

Yusuf S. (2004). Randomised clinical trials: Slow death by a thousand unnecessary policies? *Canadian Medical Association Journal*, Vol. 171 (8), pp 889–892.