Ensayos Bayesianos II: análisis de la compatibilidad frecuentista-bayesiana

Mauricio Canals Lambarri[*]

Programa de Salud Ambiental, Escuela de Salud Pública
Facultad de Medicina, Universidad de Chile

Resumen

En este artículo analizo la compatibilidad entre las aproximaciones frecuentista y Bayesiana de la inferencia estadística. Muestro primero cómo la estadística le proporciona bases al método científico, luego, cómo la estadística se basa en el concepto de probabilidad, el que tiene muchas acepciones y paradojas en su conceptualización. Enseguida estudio las bases de la estrategia frecuentista, basada en probabilidad como frecuencia relativa, y su relación con el método de máxima verosimilitud, acudiendo a los artículos originales de Fisher. Después estudio las bases de la inferencia Bayesiana y su relación con la toma de decisiones en estadística. Finalmente integro ambas aproximaciones destacando sus curiosas convergencias, para concluir que especialmente en el caso de muestras grandes, ambas aproximaciones pueden ser compatibles y convergentes.

Introducción

La estadística Bayesiana ha emergido en las últimas décadas, proponiendo un conjunto de métodos que cada día nos son más familiares en diferentes ámbitos, como epidemiología y ecología, entre otras ciencias. Algunos ejemplos son el método de Besag-York-Mollie en epidemiología y los métodos Bayesianos usados en la reconstrucción de árboles filogenéticos [1, 2]. La estadística Bayesiana se presenta corrigiendo algunos problemas epistemológicos de la estadística frecuentista [3, 4]. Parece corregir ciertos problemas, pero sin embargo, en la práctica profesional, los enfoques frecuentistas y los Bayesianos conducen a resultados muy similares. Por otra parte, durante gran parte del siglo pasado y parte de éste, la estadística frecuentista (en general paramétrica) ha sido el método fundamental en la prueba de hipótesis estadísticas en ciencias. Es decir, gran parte de la ciencia de hoy se ha apoyado en esta aproximación a pesar de sus problemas.

El objetivo de este artículo es revisar las bases y problemas de la estadística en general y los enfoques frecuentista y Bayesiano, analizando su compatibilidad: ¿existe una divergencia o convergencia entre la estadística frecuentista y la bayesiana?

El método científico

Numerosos autores contribuyeron a consolidar el llamado método hipotético-deductivo. Este tiene como pasos principales: la observación, formulación de hipótesis, deducción de consecuencias verificables, la verificación y comparación con la experiencia. René Descartes (1596-1650) definió "las reglas del método para dirigir bien la razón y buscar la verdad en las ciencias" y Francis Bacon (1561-1626) pensaba que la observación repetida de fenómenos comparables permitía extraer por inducción leyes generales que los gobiernan, siendo considerado el padre del concepto de "inducción". Sin embargo, Karl Popper (1902-1994) rechazó la posibilidad de extraer leyes generales por inducción aclarando que éstas constituyen en realidad nuevas hipótesis que permiten elaborar predicciones [5–7]. Para Popper lo central es que las teorías puedan ser refutadas ("falsabilidad") y en este sentido no habría teorías verdaderas sino no-refutadas. El método propuesto por Popper es conocido como hipotético-deductivo-refutacionista (HDR). En esta diferencia de opiniones entre Bacon y Popper parece establecerse una controversia entre la inferencia inductiva que propone la obtención de leyes generales desde la repetitividad de los fenómenos, desde lo particular a lo general (empírico, observacional), y el pensamiento deductivo que va desde lo general a lo particular (si p, entonces q). Sin embargo, de las cuatro fases del método HDR: 1) planteamiento del problema; 2) formulación de hipótesis; 3) deducción de consecuencias verificables de la hipótesis; y 4) contraste de hipótesis (refutación), la deducción sólo participa en los pasos 2 y 3. Mientras que la inducción opera en los pasos 1 y 4. Así, el método científico combina inducción y deducción. Hoy algunos autores rechazan la refutación como el único sello de lo científico [7].

[*] mcanals@uchile.cl

Estadística y la definición de probabilidad

El razonamiento inductivo conduce a la evaluación y a la toma de decisiones respecto de determinadas proposiciones, y esto se apoya en la estadística. Ésta se puede definir como "la ciencia, pura y aplicada, que crea, desarrolla y aplica técnicas para la descripción de datos y la evaluación de la incertidumbre de inferencias inductivas" (modificada de [8]). Es decir, el fin de la estadística es medir o cuantificar la incerteza de las inferencias inductivas mediante el concepto de probabilidad.

La definición clásica de probabilidad de Laplace propone que si todos los sucesos elementales del espacio muestral son mutuamente simétricos (equiprobables), la probabilidad del suceso A es el cociente entre el número de resultados favorables a A y el número de resultados posibles de un experimento:

$$\mathrm{P}(A) = \frac{\text{casos favorables a } A}{\text{casos posibles}}. \qquad (1)$$

Destacan dos observaciones: 1) en esta definición no se necesita haber realizado el experimento para conocer su probabilidad, es decir, es una definición *a priori*. En otras palabras, no sería necesario tirar una moneda para saber que la probabilidad de sacar cara es 1/2; y 2) La definición es circular, es decir, para definir probabilidad se necesita que los sucesos elementales del espacio muestral sean equiprobables. O sea, se define probabilidad en función de la equiprobabilidad del espacio muestral. Esto es conocido como la paradoja de la teoría de probabilidades.

A diferencia de este enfoque, a mediados del siglo XIX se desarrolló el concepto de probabilidad del suceso A como la frecuencia relativa de ocurrencia de A (n_A) en n intentos (n), por Antoine Agustín Cournot (1843) y Robert L. Ellis (1849). Ésta es una definición *a posteriori*, es decir, exige que el experimento se haya realizado repetidas veces antes de estimar la probabilidad de un suceso, y crea la noción frecuentista de probabilidad. Ésta fue desarrollada en mayor magnitud por Richard Von Mises (1919) definiendo la probabilidad como esta frecuencia relativa cuando se repite indefinidamente un experimento en las mismas condiciones, es decir, como un límite: $\lim_{n\to\infty} \frac{n_A}{n}$. Sin embargo, no es posible repetir un experimento infinitas veces.

Andrei Kolmogorov (1933), consciente de la paradoja de la definición clásica de probabilidad miró la probabilidad desde la teoría de la medida y propuso una definición axiomática como una función que le asigna a un evento A, un número $\mathrm{P}(A)$ entre 0 y 1, eludiendo el problema de la mutua simetría. Así, Kolmogorov considera al conjunto de los posibles resultados de un experimento como el espacio muestral (Ω), al conjunto de todos los posibles eventos como un conjunto de subconjuntos de Ω, cerrado bajo unión e intersección, es decir, un sigma-álgebra de Ω (σ-Ω) y a la probabilidad como una función

$$\begin{aligned} \mathrm{P} : \sigma\text{-}\Omega &\longmapsto [0,1] \\ A &\longmapsto \mathrm{P}(A), \end{aligned}$$

tal que

(i) $\mathrm{P}(\Omega) = 1$;

(ii) $\mathrm{P}(A) \geq 0$, para todo $A \in \sigma$-Ω;

(iii) Si $A \cap B = \emptyset$, entonces $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ (teorema de la "o").

Para Kolmogorov el triplete $(\Omega, \sigma\text{-}\Omega, \mathrm{P})$ constituye el espacio de probabilidades. Su definición es *a posteriori* ya que para determinar la medida de la probabilidad se necesita medir el evento A y tener una medida del espacio muestral [9–11].

De estas definiciones se puede observar una dualidad que se refleja en la siguiente pregunta: ¿es la probabilidad un atributo del evento, o es un atributo del conocimiento? Las definiciones *a posteriori* requieren que el experimento se haya realizado, siendo entonces la probabilidad un atributo del evento. Pero si el experimento ya se realizó y conocemos el espacio muestral y su medida, ¿para qué necesitamos el concepto de probabilidad? La definición clásica en cambio no requiere que el experimento se haya realizado, pero para su estimación se requiere el conocimiento de los sucesos elementales que constituyen el espacio muestral. Esta dualidad se puede hacer evidente en el siguiente ejemplo. Hay dos sujetos; el sujeto 1 tira una moneda al aire, ve el resultado sin que el otro lo vea y le pregunta al sujeto 2: ¿cuál es la probabilidad que sea sello? El sujeto 2 dice inmediatamente: 1/2. Pero para el primer sujeto el evento ya ocurrió y entonces este dice: ¡no! La probabilidad es 1, salió sello. El sujeto 1 conoce el evento y su frecuencia relativa es 1 (1 en 1 intento) mientras que el segundo mide su conocimiento del posible resultado, uno de dos posibles (1/2).

Estrategia de la estadística frecuentista

Gran parte de la ciencia se basa en datos y el rechazo o aceptación de hipótesis en estadística requiere la medición de la incerteza de las inferencias inductivas (p-valor). La estrategia de la estadística frecuentista consiste en: 1) Planteamiento de una hipótesis; 2) Dicotomía de la hipótesis: H_0 vs H_1, donde H_0 es la hipótesis de nulidad y H_1 es la hipótesis alternativa (de investigación); 3) Elección del nivel de significación (α). En general en ciencias se consideran como adecuados niveles de α de 0,1, 0,05 o 0,01. El más utilizado es $\alpha = 0{,}05$; 4) Elección de un estadístico (o

test) y una dócima apropiada, basados en ciertos supuestos; y 5) En base al resultado de la dócima, tomar una decisión estadística. De esta manera, la estrategia de la estadística frecuentista va dirigida al rechazo de la hipótesis de nulidad H_0, lo que permite, dado la dicotomía de la hipótesis, aceptar por defecto la hipótesis de investigación contenida en H_1. Actualmente, para la toma de decisiones se usa una estrategia combinada inicialmente desarrollada por Ronald Fisher y modificada por Jerzy Neyman y Egon Pearson. La estrategia es controlar la probabilidad de rechazar H_0 dado que ésta es verdadera, que se estima a partir de los datos: $\mathrm{P}(\text{Rechazo } H_0 \mid H_0 \text{ verdadera}) = \mathrm{P}(\text{Error de tipo I})$. Si ésta es menor que el máximo nivel que estamos dispuestos a tolerar (i.e. α), entonces rechazamos H_0 y en consecuencia aceptamos H_1. Observamos dos hechos muy curiosos: 1) lo que medimos es en esencia la probabilidad de tener los datos observados ($\mathbf{x}$) bajo el supuesto que H_0 es verdadera, $\mathrm{P}(\mathbf{x} \mid H_0)$, y no $\mathrm{P}(H_0 \mid \mathbf{x})$; si esta probabilidad es baja entonces suponemos que H_0 es falsa; y 2) nunca probamos H_1 ni medimos su probabilidad, es decir, la aceptamos por defecto al rechazar H_0. Si aceptamos H_0 (en rigor, si no la rechazamos) podemos estar cometiendo otro error (de tipo II), que es medible: $\mathrm{P}(\text{Aceptar } H_0 \mid H_0 \text{ falsa}) = \mathrm{P}(\text{Aceptar } H_0 \mid H_1 \text{ verdadera})$, pero no siempre lo medimos [8].

Para profundizar y entender mejor la inferencia frecuentista es necesario comprender algunos conceptos previos:

Población: Conjunto de objetos o eventos de interés asociados a un estudio o experimento.

Muestra: Es un subconjunto de la población. Está constituida por las unidades de muestreo, por ejemplo, "individuos". La unidad de análisis puede ser diferente de la unidad de muestreo, por ejemplo, el individuo (unidad de análisis) perteneciente a una vivienda (unidad de muestreo).

Probabilidad: es la frecuencia relativa de un evento.

Parámetro o estimando ($\boldsymbol{\theta}$): Son características de la población, por ejemplo la esperanza o promedio poblacional μ, una proporción poblacional P, la desviación estándar σ o la varianza σ^2.

Estadístico: es una función de una variable aleatoria construida a partir de la muestra, por ejemplo t.

Estadístico suficiente: Uno o más estadísticos son suficientes si no se puede calcular otro estadístico desde la misma muestra que proporcione cualquier información adicional en relación al parámetro estimado.

Estimador ($\hat{\boldsymbol{\theta}}$): es un estadístico cuya finalidad es generar un valor que se aproxime al valor de $\boldsymbol{\theta}$ a partir de las unidades de la muestra. Los ejemplos más característicos son el promedio muestral $\bar{x}$ para estimar μ, la proporción muestral $\hat{P}$ para estimar P, la desviación estándar muestral s para estimar σ y la varianza muestral s^2 para estimar σ^2.

Estimación: es el valor muestral obtenido con un estimador que aproxima el valor de un parámetro, por ejemplo $\bar{x} = 5$.

La inferencia frecuentista tiene dos capítulos muy relacionados: la estimación de parámetros y la docimasia de hipótesis. Ambas se refieren a parámetros que especifican o caracterizan la distribución de una o varias variables. Es común trabajar con distribuciones donde los parámetros ($\boldsymbol{\theta}$) se pueden particionar en el o los parámetros de interés ($\boldsymbol{\Psi}$) y uno o más parámetros de dispersión o ruido (*nuisance*, $\boldsymbol{\Lambda}$). Un ejemplo habitual es la distribución normal o Gaussiana, donde $\boldsymbol{\theta} = (\mu, \sigma)$, siendo μ el parámetro de interés y σ el de dispersión [12, 13]. En ambos capítulos de la inferencia se usa un estadístico o *pivot* que especifica el área alrededor del parámetro ψ que proporciona una medida de incerteza de éste. Por ejemplo, si nuestra hipótesis o estimación es referida a ψ (las hipótesis son siempre referidas a parámetros), entonces la idea es contar con un estadístico $T(t;\psi)$, donde t es una variable aleatoria, que permita estimar la probabilidad $\mathrm{P}(T(t;\psi) \leq t_c)$ (o equivalentemente $\mathrm{P}(T(t;\psi) > t_c)$), donde t_c es un valor crítico, con $0 < c < 1$. Esto implica que $\mathrm{P}(\psi \leq q(t,c)) = 1 - c$ (de aquí el nombre de *pivot*). Por ejemplo, en el caso del promedio μ con varianza conocida σ^2, construimos $Z = (X - \mu)/(\sigma/\sqrt{n})$ y estimamos que por ejemplo $\mathrm{P}(-1{,}96 \leq Z \leq 1{,}96) = 0{,}95$ o equivalentemente $\mathrm{P}(Z < -1{,}96 \vee Z > 1{,}96) = 1 - 0{,}95 = 0{,}05$. Además, como $\mathrm{P}(-1{,}96 \leq (X - \mu)/(\sigma/\sqrt{n}) \leq 1{,}96) = 0{,}95$, obtenemos $\mathrm{P}(X - 1{,}96\sigma/\sqrt{n} \leq \mu \leq X + 1{,}96\sigma/\sqrt{n}) = 0{,}95$, lo que llamamos intervalo del 95 % de confianza para el parámetro μ cuando σ^2 se conoce. Notemos que el valor 1,96 en este caso fue sólo un ejemplo y que en realidad corresponde a un valor de Z crítico ($z_{\alpha/2}$) para una cierta probabilidad predeterminada, en este ejemplo, $p = 1 - \alpha/2 = 0{,}95$. También en este ejemplo, en el caso de la docimasia de una cierta H_0 podemos calcular $\mathrm{P}(Z = (X - \mu)/(\sigma/\sqrt{n}) > 1{,}96 = t_c)$.

Según Cox (2006) [12] podemos encontrar dos enfoques o estrategias en la inferencia frecuentista: la estrategia (o reducción) de Fisher y el criterio operacional de Neyman-Pearson.

La estrategia de Fisher

En la estrategia de Fisher los pasos fundamentales son: a) determinar la función de verosimilitud; b) reducir a un estadístico suficiente S de la misma dimensión que θ; c) encontrar la función de S que tiene una distribución que sólo depende de ψ; y d) encontrar los cuantiles de esta distribución para obtener los límites para ψ, para un determinado nivel de probabilidad [12, 13]. Así esencialmente, la reducción de Fisher está diseñada para encontrar un estadístico suficiente para determinar el rango de resultados donde ψ puede

ocurrir en una distribución de probabilidad que define todos los valores posibles de éste parámetro. Como vemos, hay algunas claves en la estrategia de Fisher.

La primera cuestión clave incluye el concepto de verosimilitud y de la función de verosimilitud, donde verosimilitud no es lo mismo que probabilidad sino que es proporcional a ella. En palabras textuales de Fisher (1922) [14]: *"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed"* (La verosimilitud de que cualquier parámetro (o conjunto de parámetros) deba tener algún valor particular (o conjunto de valores) es proporcional a la probabilidad de que, si así fuera, la totalidad de las observaciones tengan los mismos valores que los observados). Así, la función de verosimilitud (*likelihood function*) es la densidad conjunta de los datos $f(\mathbf{x};\theta)$ pero con una pequeña diferencia. La $f(\mathbf{x};\theta)$ considera los datos $\mathbf{x}$ como argumentos y θ es fijo, mientras que la función de verosimilitud $L(\theta;\mathbf{x})$ considera los datos $\mathbf{x}$ fijos y θ es el argumento. Entonces es claro que $L(\theta;\mathbf{x})$ no especifica y es diferente de la probabilidad de un cierto valor del parámetro θ dados los datos.

La segunda cuestión clave es encontrar un estadístico suficiente para el parámetro al que se refiere la hipótesis en el caso de docimasia. Y aquí la cosa se pone complicada. Entre los miles de estadísticos que uno pudiera inventar, ¿cómo saber que el propuesto es suficiente? (i.e. no hay otro mejor que aporte nueva información de los parámetros). Fisher construye entre 1912 y 1922 un método que, según él, conduce a la obtención de estadísticos suficientes y en 1922 lo bautiza con el nombre de "método de máxima verosimilitud" (*maximum likelihood method*) [15]. Sin embargo, en este artículo expresa: *"For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of the maximum likelihood leading in any case to an insufficient statistic"* (Para la solución de problemas de estimación necesitamos un método que para cada problema particular nos lleve automáticamente al estadístico mediante el cual se satisface el criterio de suficiencia. Creo que tal método lo proporciona el método de máxima verosimilitud, aunque no estoy satisfecho en cuanto al rigor matemático de cualquier prueba que pueda presentar en ese sentido. Se invita a los lectores de las páginas siguientes a formarse su propia opinión sobre la posibilidad de que el método de máxima verosimilitud conduzca en cualquier caso a un estadístico insuficiente).

Fisher (1922) explica el método de máxima verosimilitud de la siguiente forma: "si en una distribución que involucra parámetros desconocidos $\theta_1, \theta_2, \ldots$ la probabilidad de que una observación caiga en un rango dx es $f(x;\theta_1,\theta_2,\ldots)dx$, entonces la chance (curiosamente Fisher usa esta palabra, la que hoy para nosotros tiene un significado diferente) que en una muestra de tamaño n, n_1 caigan en un rango dx_1, n_2 en un rango dx_2 y así sucesivamente, será:

$$\frac{n!}{\prod n_p!}\prod[f(x_p;\theta_1,\theta_2,\ldots)dx_p]^p\text{''}.$$

El método de máxima verosimilitud consiste en encontrar el conjunto de parámetros $(\theta_1,\theta_2,\ldots)$ que maximiza esta función, y como sólo la función f incluye los parámetros, entonces basta con maximizar $S\log f$ (i.e. el logaritmo de la parte que incluye los parámetros). Fisher lo ejemplifica con una distribución binomial, n ensayos de Bernoulli con probabilidad de éxito p, con x éxitos e y fracasos ($x+y=n$), en donde

$$\mathrm{P}(X=x)=\frac{n!}{x!y!}p^x(1-p)^y \tag{2}$$

representa la probabilidad de distribución conjunta de los datos. Entonces

$$S\log f = x\log p + y\log(1-p). \tag{3}$$

Derivando esta expresión respecto a p, e igualando a 0, obtenemos el valor $\hat{p}=x/n$, quien sería el estimador maximoverosímil y suficiente de p.

Es bien conocido que a pesar de las dudas de Fisher en relación a la rigurosidad matemática, su método es hoy ampliamente utilizado para obtener estimadores.

Entonces en la estrategia de Fisher, en el caso de la docimasia de hipótesis, tenemos una hipótesis de nulidad, H_0, que especifica numéricamente la distribución de los datos. Se desea entonces examinar la coherencia de esta hipótesis nula con los datos observados. Además, suponemos un estadístico suficiente, T (variable aleatoria), tal que cuanto mayor sea su valor observado t, más fuerte será la discrepancia en cuestión y tal que se conozca la distribución de la variable aleatoria T bajo H_0. Para evaluar la consistencia con H_0, se tiene un valor observado t, una distribución de probabilidad para T si H_0 fuera verdadera y la especificación de que cuanto mayor sea t, peor será la consistencia. Entonces no hay otra opción en esta formulación que usar $\mathrm{P}(t)=\mathrm{P}(T>t;H_0)$, que conocemos como el p-valor (*p-value*).

En la estrategia de Fisher sólo existe una posbilidad de error, cuya probabilidad está medida por el p-valor: rechazar una hipótesis de nulidad cuando es verdadera. De hecho Fisher nunca lo llamó error de tipo I, sino que propuso sólo medir el p-valor y en base a éste tomar una decisión en relación a H_0. Esta idea recuerda mucho a Popper y aunque fueron contemporáneos

(R.A. Fisher 1890-1962 y K. Popper 1902-1994), al menos la historia no registra que éstos se hayan leído mutuamente. Invito al lector a estimar la probabilidad que ambos se hayan leído, que sin duda será alta.

Criterio operacional de Neyman-Pearson

La segunda estrategia fue introducida por J. Neyman y E. Pearson, llamando al error de Fisher error de tipo I y definiendo un segundo error, error de tipo II. El nivel de significación equivale a la magnitud máxima del riesgo que está dispuesto a correr el investigador, de cometer el error de rechazar una hipótesis de nulidad cuando es verdadera. Sin embargo en la perspectiva de Neyman y Pearson, para establecer el nivel de significación estadística habría que estudiar cada tipo de error, y a partir de ahí se decidiría cuál de ellos es preferible minimizar. A partir de este último tipo de error, introdujeron el concepto de "poder o potencia de una prueba estadística", el cual se refiere a su capacidad para evitar el error tipo II, y está definido por $1-\beta$ (β es la tasa del error de tipo II). A partir de éste se ha desarrollado el concepto de "tamaño del efecto" que algunos han propuesto como sustituto de los valores p en los informes de investigación científica [16,17]. Así, Neyman y Pearson nos invitan a mirar H_0 y H_1. Aquí es útil aclarar que el p-valor no mide el tamaño del efecto ni estima la verosimilitud de H_1, lo que hace que la frase "fue estadísticamente significativo" induzca a errores de interpretación, por lo que la Sociedad Americana de Estadística (ASA) ha sugerido evitarla [18] (ver Figura 1).

Vemos que gran parte del conocimiento científico se basa en un método que tiene varias grietas conceptuales como la paradoja de las probabilidades y una estrategia al menos extraña, pero que sin embargo aceptamos como una especie de híbrido de los conceptos de Fisher, Neyman y Pearson (un estimado colega llama a esta estrategia "el imbunche estadístico", comparándola con una entidad deforme y espantosa de la mitología mapuche y chilota).

Inferencia Bayesiana

Lo que llamamos hoy teorema de Bayes [19] se puede enunciar de la siguiente forma: "Dada una partición del espacio muestral C_j, con $j = 1, 2, \ldots, k$ (i.e. causas) y un evento A (i.e. un hecho), entonces la probabilidad condicional (o *a posteriori*) de C_i, dado el evento A (i.e. la probabilidad de una causa dado un hecho) es

$$\mathrm{P}(C_i \mid A) = \frac{\mathrm{P}(A \mid C_i)\mathrm{P}(C_i)}{\sum_j \mathrm{P}(A \mid C_j)\mathrm{P}(C_j)}. \qquad (4)$$

Esto nos indica que para calcular (o actualizar) la probabilidad de un evento es necesario estimar la probabilidad *a priori* de C_i ($\mathrm{P}(C_i)$) y la probabilidad de que ocurra el evento A dado cada uno de los posibles C_j. Al extender este teorema a variables aleatorias continuas y a la estimación de parámetros (θ) a partir de los datos ($\mathbf{x}$), podemos establecer que la probabilidad o distribución *a posteriori*, dado el conjunto de datos ($\pi(\theta \mid \mathbf{x})$) es proporcional a la verosimilitud de los datos ($\mathrm{P}(\mathbf{x} \mid \theta) = L(\theta; \mathbf{x})$) por la probabilidad o distribución a priori $\pi(\theta)$ [17]:

$$\pi(\theta \mid \mathbf{x}) \propto L(\theta; \mathbf{x})\pi(\theta). \qquad (5)$$

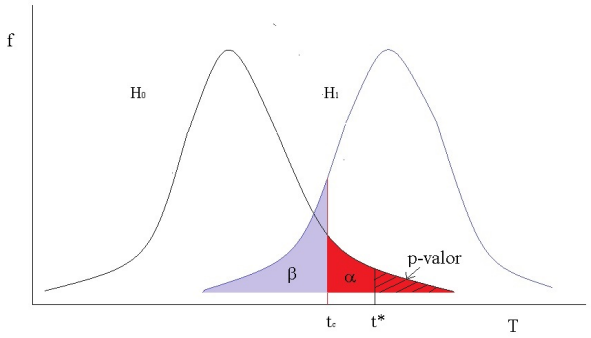

Figura 1: La estrategia estadística frecuentista. Se tienen las distribuciones de frecuencia o densidades f de dos hipótesis especificadas por sus parámetros en función de un estadístico suficiente T con un valor crítico t_c. La probabilidad de rechazo de H_0 cuando ésta es verdadera para este valor crítico arbitrario es $\alpha = \mathrm{P}(\text{rechazo de } H_0 \mid H_0 \text{ verdadera})$ (en rojo). En color gris se muestra el error de tipo II (β). El área achurada representa el p-valor determinado por el valor calculado t^* a partir de la muestra. Observamos que la idea es que el p-valor (área achurada) sea menor que el máximo valor que estamos dispuestos a aceptar representado por el nivel de significación α, en rojo. Por otra parte, al mover t_c podemos disminuir la tasa de error tipo I, pero con el costo de aumentar la de error tipo II, y viceversa.

La inferencia Bayesiana en general se realiza por métodos iterativos que permiten actualizar la probabilidad de una hipótesis dado los datos (Figura 2). Tiene la particularidad de que necesita una estimación *a priori* de la distribución de los parámetros (o las hipótesis especificadas por ellos) que puede ser subjetiva (en lo sucesivo *Priori*). Hoy, a diferencia de los estudios iniciales de Bayes y Laplace, puede o no usarse una distribución uniforme como *Priori*. Al igual que en la estadística frecuentista, a mayor cantidad de datos menor es la incerteza de las inferencias y además cuando el tamaño muestral es grande la *Priori* tiende a ser irrelevante, es decir el peso de la decisión recae en los datos o su verosimilitud (*data dominated cases* [20] o *large samples cases* [21]). La inferencia Bayesiana usa el concepto de probabilidad como atributo del conocimiento y mide directamente la verosimilitud de la hipótesis de investigación [4] (Tabla 1).

Tabla 1: Diferencias entre la inferencia frecuentista y Bayesiana (modificada de [4]).

Concepto	Inferencia frecuentista	Inferencia Bayesiana
Probabilidad	Frecuencia relativa de un evento	Grado de certeza de conocimiento
Datos	Aleatorios, muestra representativa	Fijos, son los que hay
Parámetros	Fijos	Aleatorios
Intervalo del k % de confianza	Un intervalo que incluirá el valor del parámetro en el k % de las muestras	Un intervalo con una probabilidad k de incluir el valor real del parámetro
Conclusión	$\mathrm{P}(\mathbf{x} \mid H)$ (i.e. P(rechazo $H_0 \mid H_0$ verdadera))	$\mathrm{P}(H \mid \mathbf{x})$

Interesantemente la inferencia Bayesiana parece no seguir las ideas de Popper, sino más bien las de Francis Bacon (1620). Abraza la idea de inferir la hipótesis de investigación directamente a partir de una distribución *a priori* y de la verosimilitud. Es decir, el problema de la ciencia no se trataría de rechazar infinitas H_0's, si no de afirmar o dar soporte a hipótesis de investigación (H_1). En nuestra vida cotidiana este pensamiento Bayesiano (o Baconiano) es casi natural en la toma de decisiones. Veamos algunos ejemplos. En medicina: vemos un paciente y nos formamos una concepción *a priori* acerca de su diagnóstico, le tomamos un conjunto de exámenes (datos, verosimilitud) y proponemos un diagnóstico *a posteriori*; en investigaciones policiales: proponemos *a priori* un conjunto de posibles culpables, recabamos antecedentes (datos, verosimilitud) y proponemos un culpable *a posteriori*; en general: tenemos muchas posibles soluciones a un problema (*Priori*), realizamos estudios (datos; verosimilitud), proponemos soluciones *a posteriori*. En mi experiencia personal, como médico radiólogo, los pacientes acuden con una hipótesis diagnóstica (*Priori*), tomamos y analizamos imágenes médicas (datos: ecotomografías, radiografías, tomografías computadas, etc.) y proponemos un diagnóstico radiológico (*a posteriori*).

Compatibilidad frecuentista-Bayesiana y verosimilitud

La inferencia Bayesiana y la inferencia frecuentista son conceptualmente diferentes. La primera usa el concepto de probabilidad como un grado de conocimiento o de creencia en relación a un suceso, mientras que la inferencia frecuentista interpreta la probabilidad como una frecuencia relativa de un suceso (o el límite de ella). Es difícil conciliar ambas aproximaciones. ¿Es necesario tirar un dado para saber que la probabilidad de obtener 5 es 1/6? Para los Bayesianos no y para los frecuentistas sí, y muchas veces. En algunas ciencias en general se tiene gran cantidad de datos que permiten medir frecuencias relativas (y estimar probabilidades), como la biología o las ciencias sociales, pero en otras no y es necesario recurrir a supuestos y modelos, como la ingeniería, donde se vuelven útiles los conceptos de probabilidades *a priori* y donde no hay frecuencias relativas. ¿Cómo estimaría usted la probabilidad que dos personas se encuentren entre las 12:00 y las 13:00 horas, si cada una esperará un máximo de 15 minutos? ¿Podrá usted realizar este experimento muchas veces y medir la frecuencia relativa de encuentro?

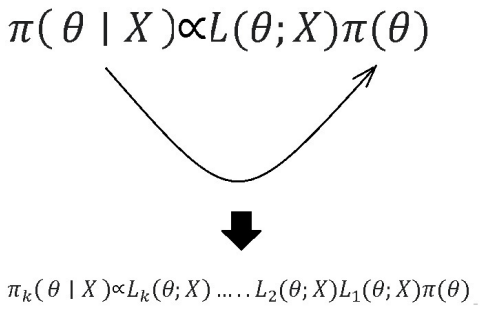

Figura 2: La estrategia Bayesiana. Se tiene una *Priori* $\pi(\theta)$ y la verosimilitud $L(\theta; \mathbf{x}) = \mathrm{P}(\mathbf{x} \mid \theta)$ que determinan una distribución *a posteriori* $\pi(\theta \mid \mathbf{x})$. Después ésta puede ser usada como una nueva *Priori* en un proceso iterativo, de manera que en la iteración k-ésima tenemos, *a posteriori*, $\pi_k(\theta \mid \mathbf{x}) \propto L_k(\theta; \mathbf{x}) \cdots L_2(\theta \mid \mathbf{x}) L_1(\theta \mid \mathbf{x})\pi(\theta)$.

Como se muestra en la Tabla 1, las diferencias no sólo se refieren al concepto de probabilidad, sino también a los datos, parámetros y a la toma de decisiones. ¿Será que una aproximación está buena y la otra mala? Y si es así, ¿cómo es posible que buena parte de la ciencia haya avanzado con éxito con una aproximación (en general, inferencia frecuentista) sin necesidad de usar la otra?

Hemos visto que en la inferencia frecuentista es fundamental encontrar estadísticos (o estimadores) suficientes que permitan caracterizar las distribuciones y sus parámetros y, en consecuencia, las hipótesis pue-

den especificarlos (por ejemplo, $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$). Hemos visto también que el método propuesto por Fisher es el método de máxima verosimilitud. Fisher tenía dudas de la matemática detrás de su método y sólo dio argumentos en favor de por qué él pensaba que su método conduce a estadísticos suficientes, dedicando todo un segmento de su artículo a justificarlo. Textualmente dice *"that the criterion of sufficiency is generally satisfied by the solution obtained by the method of maximum likelihood appears from the following considerations..."* (ver [14]).

Fisher en algunas partes de su artículo parece estar defendiendo su método contra cualquier ataque Bayesiano: *"but what a priori assumption are we to make as to the distribution of θ? Are we to assume that θ is equally likely to lie in all equal ranges $d\theta$?"*. De hecho, en su artículo de 1922 no utiliza ningún supuesto al respecto. Sin embargo, en lo que se refiere a por qué maximizar la distribución de los datos dado el parámetro (o la hipótesis) $f(\mathbf{x} \mid \theta) = L(\theta; \mathbf{x})$ y no la probabilidad del parámetro dado los datos $f(\theta \mid \mathbf{x})$, la explicación es sólo esbozada.

Fisher en 1912 [22] usa en varias ocasiones el concepto de probabilidad inversa, lo que recuerda a Laplace en 1774 [23], quien en la segunda sección de sus memorias ofrece una distinción entre aquellos casos en los que el evento (de interés) es incierto, aunque se conoce la causa, y aquellos en los que el evento es conocido y la causa es desconocida. A esto se refiere Laplace con probabilidad de las causas o probabilidad inversa. En el mismo artículo Laplace establece que "si un evento puede ser producido por un número n de causas diferentes, las probabilidades de existencia de estas causas dado el evento son como las probabilidades del evento dado estas causas" [19]. En nuestros días esto se puede expresar como [24]

$$\frac{\mathrm{P}(C_i \mid A)}{\mathrm{P}(C_j \mid A)} = \frac{\mathrm{P}(A \mid C_i)}{\mathrm{P}(A \mid C_j)}. \tag{6}$$

Esta expresión se deriva directamente del teorema de Bayes para una distribución uniforme, lo que debemos recordar para seguir la argumentación.

En 1912 Fisher [22] explica su método para estimar parámetros que entonces llamó *"an absolute criterion for fitting frequency curves"* de la siguiente forma: "Sea $p = fdx$ la chance que una observación caiga en un intervalo dx. Entonces $P = \prod fdx$ es proporcional a la chance que un conjunto dado de observaciones ocurra (distribución conjunta)". Como los factores dx son independientes de la curva teórica, *"the probability of any particular set of θ' is proportional to P, where $\log(P) = \sum \log(f)$. The most probable set of values for the θ' will make P a maximum"*. Esto lo podríamos expresar hoy como $\mathrm{P}(\theta \mid \mathbf{x}) \propto \mathrm{P}(\mathbf{x} \mid \theta)$. Éste es evidentemente un argumento basado en la probabilidad de las causas o probabilidad inversa que se encuentra en la expresión de Laplace (usted lector cambie C_i por θ y A por los datos $\mathbf{x}$ en la expresión de Laplace). En 1922 Fisher se desvincula de este argumento aunque reconoce: *"I must indeed plead guilty in my original statement of the Method of Maximum Likelihood to having based my argument upon the principle of inverse probability..."* y agrega en referencia al principio de probabilidad inversa: *"if the same observed result A might be the consequence of one or other of two hypothetical conditions X and Y, it is assumed that the probabilities of X and Y are in the same ratio as the probabilities of A occurring on the two assumptions, X is true, Y is true"*, que no es más que la expresión ya mencionada de Laplace.

Para algunos autores efectivamente Fisher se basa en la probabilidad inversa para construir el método de máxima verosimilitud. Se trata de maximizar la densidad *a posteriori* obtenida de una *Priori* uniforme [25–27]. Fisher, que según sus biógrafos, era bastante temperamental [15], en 1912 afirmó que su método sí se basaba en la probabilidad inversa, en 1917 afirmó exactamente lo mismo, en 1921 negó airadamente haber asumido una *Priori* uniforme en su trabajo de correlación, y en 1922 trató el principio por separado del postulado de *Priori* uniforme, pero reconoció que en 1912 sí lo había hecho [15]. Creo que es evidente que es así.

Así, desde este punto de vista, el método de máxima verosimilitud propone que a través de $f(\mathbf{x} \mid \theta)$ estamos mirando $f(\theta \mid \mathbf{x})$, y calculando los parámetros que optimizan $f(\mathbf{x} \mid \theta) = L(\theta; \mathbf{x})$ obtenemos estadísticos y estimadores suficientes que nos permiten especificar y probar hipótesis en la inferencia frecuentista. Esto es muy curioso, ya que le otorga una base "Bayesiana" a la inferencia frecuentista de Fisher.

Mirando ahora estos aspectos desde la perspectiva Bayesiana, la base se encuentra en el siguiente desarrollo: El teorema de Bayes permite medir directamente la verosimilitud de una hipótesis de investigación. Si sólo son posibles H_0 y H_1 (podríamos cambiar por parámetros θ_0 y θ_1, ya que las hipótesis se especifican con parámetros) y queremos medir directamente la probabilidad que H_1 sea verdadera habiendo obtenido un conjunto de datos $\mathbf{x}$, aplicando directamente el teorema de Bayes:

$$\mathrm{P}(H_1 \mid \mathbf{x}) = \frac{\mathrm{P}(\mathbf{x} \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(\mathbf{x} \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(\mathbf{x} \mid H_0)\mathrm{P}(H_0)} \tag{7}$$

$$\mathrm{P}(H_0 \mid \mathbf{x}) = \frac{\mathrm{P}(\mathbf{x} \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(\mathbf{x} \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(\mathbf{x} \mid H_0)\mathrm{P}(H_0)} \tag{8}$$

Dividiendo ambas expresiones obtenemos:

$$\frac{\mathrm{P}(H_1 \mid \mathbf{x})}{\mathrm{P}(H_0 \mid \mathbf{x})} = \frac{\mathrm{P}(\mathbf{x} \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(\mathbf{x} \mid H_0)\mathrm{P}(H_0)} = \frac{L(H_1; \mathbf{x})}{L(H_0; \mathbf{x})} \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)} \tag{9}$$

Lo que en general expresamos como: la *"Odds" a posteriori* (O_p) es igual al producto entre la razón de

verosimilitudes (LR) por la *Odds a priori* (O_0): $O_p = LR \cdot O_0$. En inferencia Bayesiana a LR habitualmente se le llama "Factor de Bayes" (K) para diferenciarlo del método de maximización de la razón de verosimilitudes. Conociendo la *Odds a posteriori* de H_1 podemos calcular la probabilidad

$$\mathrm{P}(H_1 \mid \mathbf{x}) = \frac{O_p}{O_p + 1}. \tag{10}$$

Por ejemplo, si el factor de Bayes es $K = 10$ y $O_0 = 3$, entonces $O_p = 10 \times 3 = 30$ y $\mathrm{P}(H_1 \mid \mathbf{x}) = 30/(30+1) = 0{,}968$. ¿Qué interesante, no? Si seguimos la línea de Bayes, bajo ciertos supuestos podríamos medir la probabilidad de H_1 dados los datos.

Entonces, en resumen, en esta aproximación reconocemos la *Odds a priori* (O_0), que no es más que el cociente entre dos *Prioris*, el factor de Bayes que es el cociente entre verosimilitudes y la *Odds a posteriori* (O_p). Notablemente, si las *Prioris* fueran uniformes $O_0 = 1$, entonces todo el peso recaería en los datos y sería un problema de verosimilitudes, tal como sugiere la estrategia frecuentista. Curiosamente también el factor de Bayes, en el caso simple de dos hipótesis (o parámetros), contiene en el denominador a $\mathrm{P}(\mathbf{x} \mid H_0) = L(H_0; \mathbf{x})$, que es justamente lo que pretende medir el p-valor en la estrategia de Fisher. Es decir, un p-valor pequeño (i.e. $p < 0{,}05$) determinará un factor de Bayes grande y, en consecuencia, una O_p y $\mathrm{P}(H_1 \mid \mathbf{x})$ grandes, apoyando favorablemente a H_1. A la inversa, con un p-valor grande, el factor de Bayes será menor y O_p y $\mathrm{P}(H_1 \mid \mathbf{x})$ pequeños, apoyando débilmente o definitivamente no apoyando a H_1. Notamos también que el factor de Bayes contiene en el numerador a $\mathrm{P}(\mathbf{x} \mid H_1) = L(H_1; \mathbf{x})$, que es análogo a la potencia de la dócima $(1-\beta)$. Es decir, justamente lo que nos proponen Neyman y Pearson con su error de tipo II. En otras palabras, Neyman y Pearson nos invitan no sólo a mirar el error de tipo I sino también mantener bajo el error de tipo II, lo que en consecuencia redundará en un factor de Bayes muy alto y en definitiva un apoyo a H_1 (y rechazo de H_0). La inferencia Bayesiana se concentra en el factor de Bayes, pero este ejemplo nos muestra que también la inferencia frecuentista lo hace, aunque de una forma diferente, sugiriendo que ambos enfoques, aunque conceptualmente diferentes, son convergentes en la toma de decisiones en estadística (Figura 3).

Conclusiones destacables

La ciencia se apoya en el método científico y éste en la estadística. Bacon propone que este método se basa en la inducción mientras que Popper propone que este método se basa en la hipótesis-deducción-refutación. La estadística frecuentista se alinea con las ideas de Popper concentrándose en la refutación de la hipótesis de nulidad, mientras que la inferencia Bayesiana parece apoyar a Bacon afirmando la hipótesis de investigación. La estadística se apoya en la medida de la probabilidad. Sin embargo, el concepto de probabilidad es primero paradojal por su propia definición circular y es diferente para las inferencias frecuentista y Bayesiana. La inferencia frecuentista trata de evitar la subjetividad Bayesiana, tratando la probabilidad como una frecuencia y buscando estadísticos suficientes para especificar y docimar hipótesis, pero en su búsqueda usa la máxima verosimilitud que tiene una base Bayesiana. La inferencia Bayesiana se basa en *Prioris* que pueden ser subjetivas o mínimamente informativas (como la distribución uniforme) y en la función de verosimilitud para determinar la distribución *a posteriori* de parámetros (o hipótesis especificadas por parámetros), pero la verosimilitud (o el factor de Bayes) es la base de la inferencia frecuentista. Finalmente, este análisis nos muestra que ambas aproximaciones son compatibles a pesar de sus diferencias conceptuales y que cuando los tamaños muestrales son grandes, el peso de las decisiones recae en los datos, y ambas aproximaciones convergen.

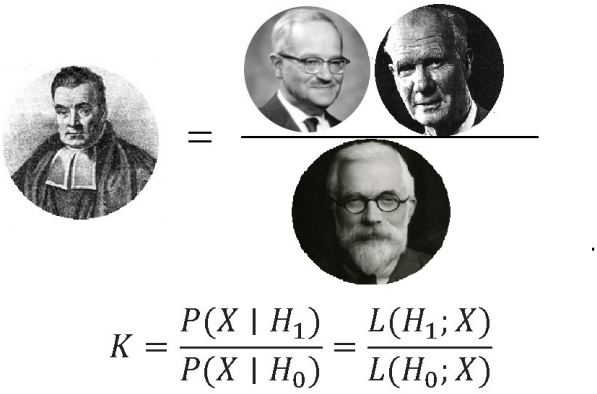

Figura 3: El factor de Bayes (K) es el cociente entre verosimilitudes para dos diferentes hipótesis especificadas por sus parámetros. En el denominador se encuentra la idea de un conservador Fisher cuya estrategia propone sólo un p-valor bajo $\mathrm{P}(\mathbf{x} \mid H_0)$. Jerzy Neyman y Egon Pearson nos invitan también a mirar en numerador $\mathrm{P}(\mathbf{x} \mid H_1)$, mientras que la idea que usa la inferencia Bayesiana es el cociente K, que en definitiva determina O_p y $\mathrm{P}(H_1 \mid \mathbf{x})$.

Referencias

[1] Canals M, Canals A, Ayala S, Valdebenito J, Alvarado S, Cáceres D. (2020). Changes in age and geographic distribution of the risk of Chagas disease in Chile from 1989 to 2017. *Vector Borne and Zoonotic Diseases*, *21*(2), 98-104. doi.org/10.1089/vbz.2020.2647

[2] Nascimento FF, dos Reis M, Yang Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecol & Evol*, 1:1446-1454.

[3] Bernardo J, Smith AFM. (1994). *Bayesian Theory*. Hoboken, New Jersey: John Wiley & Sons, Inc.

[4] Ellison AM. (1996). An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol Appl*, 6:1036-46.

[5] Tankard JW. (1984). *The Statistical Pioneers*. Cambridge, MA: Schenkman Publishing Co.

[6] Rodríguez E. (2005). Estadística y Psicología: Análisis histórico de la inferencia estadística. *Perspectivas psicológicas*, 15:96-102.

[7] Bunge M. (2010). *Las pseudociencias*. Pamplona: Editorial Laetoli.

[8] Steel RGD, Torrie JH. (1980). *Bioestadística: Principios Y Procedimientos*. Bogotá: McGraw-Hill Latinoamericana S.A.

[9] Kolmogorov A. (1956). *Foundations of the theory of probability*, 2nd Ed. Dover Pub.

[10] Feller W. (1968). *An Introduction To Probability Theory And Its Applications*. New York: John Wiley & Sons.

[11] Cramer H. (1996). *Elementos de la Teoría de Probabilidades*. Madrid: Aguilar S.A.

[12] Cox, DR. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

[13] Reid N, Cox DR. (2014). On Some Principles of Statistical Inference. *International Statistical Review*, *83*(2):293-308. 10.1111/insr.12067.

[14] Fisher RA. (1922). On the mathematical foundations of theoretical statistics. *Phill Trans Roy Soc London A*, 222:309-368.

[15] Aldrich J. (1997). R.A. Fisher and the Making of Maximum Likelihood 1912–1922. *Statistical Science*, *12*(3):162-176.

[16] Murphy KR, Myors B. (2004). *Statistical Power Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

[17] Canals M. (2019). Bases científicas del razonamiento clínico: inferencia Bayesiana. *Rev Med Chile*, 147:231-237.

[18] Wasserstein RL, Lazar NA. (2016). ASA statement on statistical significance and p-values. *Amer Stat 70*(2):12933.

[19] Canals M. (2023). Ensayos Bayesianos I: buscando a Bayes. *Inferencias*, 9:14-20.

[20] Box GEP, Tiao GG. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

[21] Kass RE, Wasserman L. (1996). The selection of prior distributions by formal rules. *Amer Stat Asoc 91*(435):1343-1370.

[22] Fisher RA. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155-160.

[23] Laplace PS. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des Sciences de MI (Savants étrangers)* 4:621-656. Reimpreso en: Laplace, *Oeuvres complètes* (París, Francia: Gauthier-Villars et fils, 1841), vol.8, pp.27-65.

[24] Dale AI. (1982). Bayes or Laplace? An examination of the origin and early applications of Bayes´ theorem. *Archive for History of Exact Sciences*, *27*(1):23-47.

[25] Mackenzie DA. (1981). *Statistics in Britain 1865-1930*. Edinburgh Univ. Press.

[26] Geisser SG. (1980). *Basic theory of the 1922 mathematical statistics paper*. In *R.A. Fisher: An Appreciation*, Fienberg SE, Hinkley DV, eds, New York: Springer, pp:59-66.

[27] Conniffe DC. (1992). Keynes on probability and statistical inference and the links to Fisher. *Cambridge Journal of Economics*, 16:475-489.