

## **$p < 0,05$ , ¿Criterio mágico para resolver cualquier problema o leyenda urbana?**

**Pedro Monterrey Gutiérrez**

Departamento de Matemáticas, Universidad del Rosario, Bogotá, D.C., Colombia.

[pedro.monterrey@urosario.edu.co](mailto:pedro.monterrey@urosario.edu.co)

Recibido: 08-05-2012; Aceptado: 03-07-2012

### **Resumen**

Las Pruebas de Hipótesis son el procedimiento de análisis más conocido por los investigadores y utilizado en las revistas científicas pero, a su vez, ellas han sido fuertemente criticadas, su uso ha sido cuestionado y restringido en algunos casos por las inconsistencias observadas en su aplicación. Este problema se analiza, en este artículo, tomando como punto de partida los Fundamentos de la Metodología Estadística y los diferentes enfoques que históricamente se han desarrollado para abordar el problema del análisis de las Hipótesis Estadísticas. Resaltándose un punto poco conocido por algunos: el carácter aleatorio de los valores P. Se presentan los fundamentos de las soluciones de Fisher, Neyman-Pearson y Bayesiana y a partir de ellas se identifican las inconsistencias del procedimiento de conducta que indica identificar un valor P, compararlo con el valor del error de tipo I –que usualmente es considerado como 0,05– y a partir de ahí decidir las conclusiones del análisis. Adicionalmente se identifican recomendaciones sobre cómo proceder en un problema, así como los retos a enfrentar, en lo docente y en lo metodológico, para analizar correctamente los datos y determinar la validez de las hipótesis de interés.

**Palabras clave:** Pruebas de Hipótesis de Neyman-Pearson. Pruebas de Significación de Fisher. Pruebas de Hipótesis Bayesianas. Normas de Vancouver. Valores P. Hipótesis nada.

### **Abstract**

**$p < 0.05$ , A magic criterion to solve any problem or an urban legend?** Hypothesis testing is a well-known procedure for data analysis widely used in scientific papers but, at the same time, strongly criticized and its use questioned and restricted in some cases due to inconsistencies observed from their application. This issue is analyzed in this paper on the basis of the fundamentals of the statistical methodology and the different approaches that have been historically developed to solve the problem of statistical hypothesis analysis highlighting a not well known point: the P value is a random variable. The fundamentals of Fisher's, Neyman-Pearson's and Bayesian's solutions are analyzed and based on them, the inconsistency of the commonly used procedure of determining a p value, compare it to a type I error value (usually 0.05) and get a conclusion is discussed and, on their basis, inconsistencies of the data analysis procedure are identified, procedure consisting in the identification of a P value, the comparison of the P-value with a type-I error value –which is usually considered to be 0.05– and upon this the decision on the conclusions of the analysis. Additionally, recommendations on the best way to proceed when solving a problem are presented, as well as the methodological and teaching challenges to be faced when analyzing correctly the data and determining the validity of the hypotheses.

**Key words:** Neyman-Pearson's hypothesis tests, Fisher's significance tests, Bayesian hypothesis tests, Vancouver norms, P-value, null-hypothesis.

## Resumo

**p < 0,05, Critério mágico para resolver qualquer problema ou lenda urbana?** Os testes de hipóteses são o método de análise melhor conhecido por pesquisadores e utilizado em revistas científicas; mas por sua vez, têm sido fortemente criticados, seu uso tem sido questionado e, em alguns casos restritos pelas inconsistências observadas na sua aplicação. Esse problema é discutido neste artigo, tendo como ponto de partida os Fundamentos da Metodologia Estatística e as diferentes abordagens que historicamente têm sido desenvolvidas para resolver o problema da análise das Hipóteses Estatísticas. Destacando-se um ponto pouco conhecido por alguns: o caráter aleatório do p-valor. Apresentam-se os fundamentos das soluções de Fisher, Neyman-Pearson e Bayesiana e delas são identificadas as inconsistências do procedimento de conduta que orienta identificar um p-valor para compará-lo com o valor do erro de tipo I, que é geralmente considerado como 0,05 - e, posteriormente, decidir as conclusões da análise. Além disso, se identificam recomendações sobre como proceder num problema, e os desafios a serem enfrentados no ensino e no metodológico, para analisar corretamente os dados e determinar a validade das hipóteses de interesse.

**Palavras-chave:** teste de hipóteses de Neyman-Pearson, testes de significância de Fisher, testes de hipóteses Bayesianas, normas de Vancouver, p-valor, hipótese nula.

## Introducción

Según la American Statistical Association “La Estadística es la aplicación científica de principios matemáticos a la recolección, análisis y presentación de datos numéricos” (1). En el campo de la Investigación Cuantitativa esto se refleja en que la Estadística constituye un elemento metodológico fundamental pues, en este tipo de investigación, las evidencias se derivan de la obtención y análisis de información (2). En ocasiones el trabajo estadístico en la investigación de las diferentes ciencias es realizado total o parcialmente por los propios investigadores, quedando relegada la participación de los profesionales del área a algunos momentos puntuales dentro de las diferentes etapas del trabajo estadístico, lo que en múltiples ocasiones se reduce al análisis de los datos. Esta situación no es muy estimulante pues, como resultado de los énfasis que se hace en los programas de formación en estadística para las diferentes ciencias, así como de sus falencias, muchos investigadores desestiman, por no decir desconocen, los retos metodológicos de la etapas de muestreo y en la selección de las mediciones, así como su efecto en los criterios para el análisis de los datos. En múltiples ocasiones el investigador “selecciona” una muestra de individuos con criterios ecléticos y organiza la investigación como un ejercicio metodológico cuyo fin es coleccionar observaciones para realizar con ellas, posteriormente, Pruebas de Hipótesis con el objetivo de determinar “la significación de los resultados” y con eso dar respuesta a la pregunta de investigación. En general no es costumbre introducir en los análisis las peculiaridades de las mediciones en estudio ni las propiedades de los datos que se pudieran derivar del criterio de muestreo empleado, siendo analizadas las observaciones, en un elevadísimo porcentaje de los casos, con las mismas pruebas de hipótesis independientemente de sus características específicas. De esta forma las Pruebas de Hipótesis se han convertido, en mayor o menor medida en el fin de la investigación, su terminología de “diferencias significativas” ha permeado la investigación y el lenguaje

diario. La búsqueda de un valor p y su comparación con el valor del error de tipo I, que en lo general es fijado como 0,05, se ha convertido en un algoritmo generador de las conclusiones de la investigación. Una mirada a cualquier revista científica mostrará el predominio de las pruebas de hipótesis como criterio central del análisis de los datos que se presentan y la fe absoluta en el criterio  $p < 0,05$  como único generador de las conclusiones de muchas investigaciones.

A pesar de su popularidad es un hecho que las Pruebas de Hipótesis, como metodología, se encuentran en un momento crítico; ellas han sido cuestionadas como las causantes de inconsistencias en la investigación, llegando al extremo de que, por ejemplo, las normas de Vancouver, normas que rigen la estructura que deben tener las publicaciones biomédicas, promuevan el que no sean utilizadas (3) y a que publicaciones de diferentes áreas desestimen su uso, por ejemplo, la revista *Epidemiology*, en sus instrucciones a los autores dice “... desestimamos enfáticamente el uso de valores P y el lenguaje referido a la significación estadística...” (4). En múltiples esferas del conocimiento, donde se aplica la estadística en sus investigaciones, existe una fuerte discusión acerca de la pertinencia o no del uso de las Pruebas de Hipótesis. Lamentablemente estas discusiones e incertidumbres no son del conocimiento de muchos investigadores y en ocasiones de algunos profesionales de la Estadística. Si bien es cierto que en algunos casos estas controversias y críticas caen en el campo de la Filosofía de la Ciencia, en sus aspectos más comunes los problemas son consecuencia del conocimiento limitado acerca de las Pruebas de Hipótesis, sus alcances y imitaciones; así como de la no comprensión de los fundamentos del trabajo estadístico. Consejos para solucionar el problema, como sustituir las Pruebas de Hipótesis por Intervalos de Confianza (2,3), aunque portan la lógica de presentar la medida de los efectos y el efecto del azar sobre sus estimaciones no son comprendidos por todos, lo que ha propiciado nuevas inconsistencias al desconocer aspectos conceptuales e históricos de ambos contenidos.

El objetivo de este artículo es presentar cómo se concibe el trabajo estadístico en el tema de las Pruebas de Hipótesis en la actualidad, mostrando cómo los libros de texto en uso no han logrado la actualidad que se espera de ellos y siguen presentado la estadística con los mismos paradigmas de mediados del siglo pasado. En el campo de las propias pruebas se presentará cuáles son las raíces metodológicas de los problemas en su aplicación, desmitificando algunos paradigmas muy arraigados entre quienes las usan, mostrando sus alcances y limitaciones, así como las vías para proceder frente a los problemas de análisis de las hipótesis estadísticas.

### **La metodología estadística. Sus fundamentos**

La recolección de los datos es la fase inicial del trabajo estadístico. En este aspecto la identificación de estrategias de muestreo, en correspondencia con el problema a analizar, constituye un elemento fundamental para que los datos cumplan con su función principal que es, brindar información que permita obtener respuestas a las interrogantes del problema. La importancia de este tema fue abordada en 1987 por CR. Rao, discípulo de RA Fisher y uno de los estadísticos más destacados del pasado siglo, cuando dijo que el Muestreo y el Diseño Experimental eran las dos metodologías más importantes de la Estadística. La posición de Rao fue fundamentada al afirmar que "...si los datos son buenos, los resultados deberían resultar obvios..." (5). La conveniencia de prestar una relevante atención a los esquemas de muestreo radica, entre otros elementos, en que muestras tomadas sin una estrategia muestral, en correspondencia con el problema a abordar y las características de las poblaciones en estudio, pudieran resultar sesgadas, tal vez no ser lo suficientemente informativas o imponer restricciones relevantes al análisis de datos, restricciones que, en ocasiones, no son apreciadas por quién la selecciona y analiza posteriormente.

Una vez tomados los datos procede realizar su análisis con la intención de obtener de ellos la información buscada; pero, contrario a lo que algunos pudieran pensar, la realización de una Prueba de Hipótesis no es el punto central del análisis, ni la única forma de hacerlo. Hasta principios del siglo XX los análisis de los datos se realizaban de manera muy descriptiva y solamente en contadas situaciones se ejecutaban análisis con objetivos más amplios, utilizando criterios bayesianos (6). Las limitaciones computacionales en esos tiempos restringían bastante las posibilidades de los análisis de tipo bayesiano y al no disponerse de criterios objetivos de análisis, las publicaciones científicas estaban dominadas por presentaciones anecdóticas cargadas de subjetivismo. En ese contexto surgieron las Pruebas de Hipótesis en los primeros años del siglo XX. Con el surgimiento y posterior extensión de estos procedimientos de inferencia se presentó un ligero divorcio entre la componente descriptiva y la

inferencial; las inferencias basadas en las Pruebas de Hipótesis se convirtieron, por no decir degeneraron, en un "ritual" (7) que sustituyó, en muchos casos, la capacidad de los investigadores de realizar un análisis lógico por un algoritmo de decisión donde el objetivo de los análisis era aceptar o rechazar una hipótesis. Esa etapa, de divorcio entre la teoría estadística y los problemas de aplicación, ha sido conocida como un momento en el que "...las teorías buscaban los datos en lugar de tratar con problemas reales que necesitaban de una teoría..." (8). Pasado casi un siglo de ese momento y después de múltiples críticas y llamados de alerta, la mirada acerca de cómo debe concebirse el análisis de los datos ha ido cambiando; identificándose dos criterios que pudieran constituir momentos complementarios del análisis de un problema: Los Criterios Exploratorio y Confirmativo.

El Análisis de Datos Exploratorio fue introducido por Tuckey en 1970 (9). Este análisis consiste en un conjunto de técnicas descriptivas, de tipo numérico o gráfico, que sin descansar en un modelo específico permiten identificar las evidencias en los datos y ayudan a construir hipótesis. En la actualidad estas técnicas se vinculan con procedimientos estadísticos robustos y la Estadística no Paramétrica, derivada de la Computación Estadística, con lo que las posibilidades de identificar estructuras en los datos se han potenciado. Uno de los resultados primarios del análisis exploratorio es la identificación de modelos para los datos, estos modelos, univariados o multivariados, consisten en aspectos tan primarios, como la identificación de la distribución de probabilidad que siguen los datos, hasta la estructuración de relaciones entre las variables. En la actualidad existen una gran cantidad y diversidad de técnicas descriptivas que pudieran resultar útiles en la componente exploratoria. La importancia de la fase exploratoria fue reconocida por Fisher a mediados del siglo XX; él enfatizaba que la componente gráfica era una parte muy importante en el análisis de los datos (5). Dada la potencia de estos criterios de análisis en algunos casos las evidencias mostradas en ellos son tan fuertes que resultarían innecesarios otros criterios de análisis más formales o concluyentes (8). Dicho en otras palabras no siempre es necesario aplicar las populares Pruebas de Hipótesis.

El Análisis de Datos Exploratorio de Tuckey parte de examinar la estructura de los datos e identificar, de manera primaria, relaciones entre variables; con lo que se sientan las bases para un análisis de datos más formal. En ese sentido reducir el Análisis de Datos a la fase Confirmativa produce un divorcio entre los datos y los análisis al ser esta una fase analítica. En especial, pensar en las Pruebas de Hipótesis como centro análisis de los datos, constituye una limitación para la identificación de las evidencias contenidas en ellos.

En el Análisis Confirmativo se evalúa la fortaleza de las evidencias obtenidas en la fase exploratoria, mediante la comprobación de las hipótesis, utilizando criterios probabilísticos en los que es posible establecer juicios probabilísticos acerca de su precisión. En esta fase se evalúan los modelos identificados en la fase exploratoria; en ella los procedimientos de Inferencia Estadística tienen una función relevante, aunque hay que resaltar que en la actualidad se han desarrollado procedimientos no paramétricos y semi paramétricos para ajustar modelos a los datos que permiten redimensionar los alcances de las técnicas tradicionales de inferencia.

Las Pruebas de Hipótesis se construyen para identificar en los datos evidencias a favor o en contra de las Hipótesis Estadísticas, las que guían el análisis de datos en el área de la Inferencia Estadística. Ellas no deben confundirse con la Hipótesis de Investigación. En general una Hipótesis de Investigación es “una idea o sugerencia acerca de una cierta pregunta considerada como punto de partida de los razonamientos y explicaciones del problema” (10), es una posible respuesta a la pregunta de investigación (2). Estas hipótesis científicas, generales en su naturaleza, en situaciones en que los procedimientos deductivos no son aplicables no pueden ser evaluadas directamente, por ese motivo en esos casos se debe realizar un estudio de campo o experimental para generar datos que permitan comprobar su validez. Para ello se determinan hipótesis de trabajo, las denominadas Hipótesis Estadísticas, que son hipótesis de naturaleza probabilística y guían el análisis de los datos en la dirección de identificar las evidencias a favor o en contra de la validez de la Hipótesis de Investigación.

La ciencia no deductiva parte de raíces experimentales, se construye a partir de una sucesión de evidencias; por ese motivo la razón de un estudio particular no puede ser brindar conclusiones que den solución a un problema. Su función metodológica es permitir hacer ajustes en el grado de certidumbre acerca de la validez de la hipótesis de interés. La fortaleza o debilidad de tal juicio debe ser una medida probabilística que establezca la plausibilidad o verosimilitud de tales hipótesis, en correspondencia con la naturaleza aleatoria de los estudios de los que se obtienen las evidencias. La aceptación o el rechazo de una Hipótesis de Investigación debe ser consecuencia de la acumulación de evidencias en el proceso cognitivo. El resultado final de la investigación debe ser un grado de confianza sobre un conjunto de proposiciones y esto constituye la base de las decisiones (11). Los procedimientos de análisis estadístico son solamente los criterios con los que el investigador identifica las evidencias obtenidas en un estudio particular, el reto metodológico es ubicar estas evidencias en el proceso de construcción del nuevo conocimiento, quedando determinada así la función

metodológica de la estadística y sus procedimientos en el proceso de construcción del conocimiento científico. En el caso de las Pruebas de Hipótesis, como se verá a continuación, esto ya fue establecido por los creadores de los diferentes enfoques metodológicos en los que se sustenta su análisis.

## Los diferentes enfoques para el análisis de las hipótesis estadísticas. Las raíces históricas y metodológicas de las Inconsistencias en su Uso

### Los valores P de Fisher

Alrededor de los años 20 del siglo pasado RA Fisher desarrolló un procedimiento, que él denominó Pruebas de Significación, para analizar las Hipótesis Estadísticas. Su criterio está basado en la realización de un proceso de inferencia inductiva para enjuiciar las evidencias contenidas en los datos (12). Fisher se basó, para fundamentar su criterio, en la necesidad de acumular evidencias a favor o en contra de una hipótesis como criterio de decisión acerca de su validez, para Fisher “las conclusiones a partir de su proceso de pruebas de significación son pasos con los que el investigador obtiene una mejor comprensión de su experimento” (13). Por ese motivo en el enfoque de Fisher sólo se considera una hipótesis,  $H_0$  o hipótesis nula, la que puede ser rechazada o no. Esta lógica de análisis tiene semejanzas con los planteamientos filosóficos de Karl Popper quien estableció que “la ciencia no avanza aceptando hipótesis verdaderas, sino rechazando falsas” (10). Las Pruebas de Significación de Fisher se basan en la determinación de un valor P que es una medida de la fortaleza de la evidencia de los datos en contra de la hipótesis  $H_0$ . Considerando, para ejemplificar, una dirección de rechazo unilateral, es decir identificando que los valores incompatibles con la hipótesis nula son los que se encuentran en la cola superior de la distribución, el valor P se define formalmente así: si  $\vec{\chi}$  representa el n-vector aleatorio de las observaciones,  $\Theta \subset \mathbb{R}^k$  ( $\mathbb{R}$  conjunto de los números reales) es el Espacio Paramétrico que determina la distribución de  $\vec{\chi}$  y la hipótesis estadística de interés se representa mediante  $H_0: \theta \in \Theta_0, \Theta_0 \subset \Theta$ . Si  $T_n(\vec{\chi}; \theta)$  es una transformación de los datos –estimador- que brinda información sobre  $\theta$ ; el valor P es una variable aleatoria definida por  $P = \text{Sup}_{\theta \in \Theta_0} \bar{F}_n(T_n, \theta)$  siendo  $\bar{F}_n = 1 - F_n$  y denotándose mediante  $F_n$  la función de distribución de  $T_n$  (14). La noción intuitiva que subyace en la definición del valor P es que este representa la probabilidad de los datos observados o de valores más extremos que ellos en la dirección en que contradicen la validez de  $H_0$  suponiendo  $H_0$  cierta, es decir en la cola superior de la distribución de  $T_n$  en el caso ejemplificado anteriormente. Fisher no indicó claramente cómo proceder para la identificación de la dirección en que los valores pudieran ser extremos, no planteó una hipótesis alterna que represente esta dirección,

asumió que la dirección más extrema es un concepto implícito al problema y como tal no definible formalmente. Su intención era desarrollar un procedimiento intuitivo que auxiliara en el análisis de los datos.

El procedimiento de las Pruebas de Significación de Fisher rechaza  $H_0$  cuando se obtiene en un estudio un valor P pequeño, lo que se interpreta como una evidencia en contra de  $H_0$ . Este criterio inductivo, construido a partir de los datos, se deriva de un procedimiento de análisis basado en el criterio de demostración por reducción al absurdo: El valor P se calcula, suponiendo  $H_0$  cierta, como la probabilidad de un valor muestral observado y de valores más extremos en la dirección de rechazo, si el valor P es pequeño esto tiene dos posibles interpretaciones: (a) algo poco probable ha ocurrido en este experimento o, más plausiblemente, (b) la premisa sobre la que se calculó el valor P,  $H_0$ , ha sido contradicha por los datos y en correspondencia debe ser rechazada. Para Fisher "... cada experimento se puede decir que existe para dar a los hechos la oportunidad de refutar la hipótesis nula..." (13). El centro de las Pruebas de Significación radica en la búsqueda de evidencias contra la hipótesis nula, la que se rechaza o no. Como criterio de análisis Fisher propuso utilizar flexiblemente el umbral 0,05 como medida de P pequeño, pero rechazó la posibilidad de utilizarlo como un esquema rígido de análisis y resaltó enfáticamente el papel de un estudio individual en el proceso de rechazo de una hipótesis al afirmar "...un fenómeno se considera establecido experimentalmente cuando un experimento diseñado adecuadamente raras veces deja de dar este nivel de significación...". De esta forma las conclusiones científicas se derivan de un proceso inductivo en el que el conocimiento se va construyendo gradualmente como suma de evidencias experimentales.

La naturaleza aleatoria del valor P, el que es una variable aleatoria al ser derivado de los datos, refuerza el papel parcial de las evidencias obtenidas en un estudio individual. Es fácil demostrar que el valor P, si  $H_0$  es cierta, sigue una Distribución Uniforme con valores concentrados en el intervalo  $[0;1]$ , esto significa que en el caso en que  $H_0$  sea cierta los valores P obtenidos en estudios semejantes pueden encontrarse con probabilidades constantes en el rango  $0 \leq p \leq 1$ ; en particular esta afirmación se traduce en que, si  $H_0$  es cierta, en el 5% de los estudios que se realicen se obtendrá un valor  $p < 0,05$  sólo como consecuencia del azar. Más difícil es el análisis de la distribución de probabilidad de los valores P cuando  $H_0$  no es cierta, esta distribución, ciertamente, depende de la dirección de rechazo, sin embargo la determinación de la distribución de P en este caso es muy importante para la realización de los

estudios de integración de evidencias de diferentes estudios o meta-análisis. Es un hecho que la distribución de P si  $H_0$  no es cierta es asimétrica (6), su determinación debe ser casuística, sin embargo, Lambert y Hall (14) demostraron que bajo condiciones bastante generales esta distribución es asintóticamente log normal. Hung y colaboradores (15) analizaron esa distribución en algunos casos especiales, resaltando su vinculación con la función de potencia y el tamaño de muestra. Sackrovitz (16) obtuvo la expresión del valor esperado de P cuando  $H_0$  no es cierta, resaltando su relación con el tamaño de muestra, aunque esta relación pudiera nos ser muy útil en el caso de distribuciones muy asimétricas.

Para el uso e interpretación del valor P Sterne y Smith (12) propusieron rangos de valores para interpretarlo, mostrando un gradiente en los niveles de fortaleza o debilidad de la evidencia contenida en los datos contra la hipótesis nula en correspondencia con el significado de P como una medida de evidencia que necesariamente porta ciertos patrones de incertidumbre.

## Las pruebas de hipótesis de Neyman y Pearson

Las Pruebas de Significación fueron desarrolladas por Fisher para hacer inferencias a partir de un proceso inductivo que conduciría al conocimiento. Analizado como teoría matemática, su procedimiento adolecía de algunas fallas o limitaciones al ser construido sobre nociones inespecíficas. Hablando matemáticamente, en las Pruebas de Significación no había una definición clara de qué entender por "valores más extremos que los observados" ni se necesitaba una hipótesis alternativa, aspecto que aun en la actualidad genera controversias. Esta insuficiencia condujo a Jerzy Neyman y Egon Pearson en 1933 (17) a desarrollar otro criterio para abordar el problema de la validez de las Hipótesis Estadísticas, su desarrollo teórico fue denominado por ellos **Pruebas de Hipótesis**, partía de una sólida aplicación de las Matemáticas y fueron desarrolladas como una teoría matemática derivada directamente de la Teoría de Probabilidades. Por ese motivo esta solución al problema del análisis de la validez de las Hipótesis Estadísticas se identifica como un mecanismo de comportamiento inductivo, opuesto en su esencia a la inferencia inductiva de las Pruebas de Significación de Fisher. El rasgo distintivo de las Pruebas de Hipótesis de Neyman y Pearson es que se construyen considerando, junto a la hipótesis  $H_0$  de las Pruebas de Significación, otra hipótesis,  $H_A$  o alternativa, que representa la conducta a seguir en el caso en que  $H_0$  sea rechazada. En esta teoría las hipótesis representan las posibles decisiones en el problema y como tal se analizan en parejas ( $H_0, H_A$ ). En realidad Fisher reconocía

la necesidad de una alternativa como forma de identificar el criterio de valores más extremos en la dirección de rechazo se  $H_0$ , pero esta decisión quedaba a los detalles del problema pues, para Fisher, el análisis de las evidencias contenidas en los datos para rechazar o no la hipótesis era solamente un proceso intuitivo y parte de una secuencia de análisis o criterios de juicio para ampliar los conocimientos. Las críticas a la ausencia de una alternativa pueden resumirse con un comentario de Berkson “ si un evento ha ocurrido la pregunta no debe referirse a si es raro bajo  $H_0$ , la cuestión sería ver si en alguna alternativa pudiera ser relativamente frecuente ...” (6).

Las Pruebas de Hipótesis consideran el mismo punto de partida de las Pruebas de Significación:  $\vec{X}$  representa el n-vector aleatorio de las observaciones,  $\Theta \subseteq \mathcal{R}^k$  es el Espacio Paramétrico que determina la distribución de  $\vec{X}$ , la que puede ser representada por la familia de medidas de probabilidad  $(P_\theta)_{\theta \in \Theta}$ . La hipótesis estadística de interés se representa, al igual que en la formulación de Fisher, mediante  $H_0: \theta \in \Theta_0$ ,  $\Theta_0 \subset \Theta$ . A este esquema se le añade la hipótesis alternativa  $H_A: \theta \in \Theta_1$ ,  $\Theta_1 = \Theta \setminus \Theta_0$ . De nuevo se considera  $T_n(\vec{X}; \theta)$  una transformación de los datos, un estimador, que brinda información sobre  $\theta$ ;  $T_n$  se utiliza para construir una región  $K \subset \mathcal{R}^n$ , la región de rechazo, de forma tal que  $\vec{X} \in K$  identifica los datos,  $\vec{X}$ , con los que la decisión a tomar debe ser rechazar  $H_0$  y en caso contrario aceptar  $H_A$ .

Para controlar las tasas de error en las dos direcciones en que estos pudieran cometerse Neyman y Pearson introdujeron la función de potencia definida por  $\beta(\theta) = P_\theta(\{\vec{X} \in K\})$  para  $\theta \in \Theta$ . Representando respectivamente  $\beta / \Theta_0$  los errores de tipo I (rechazar  $H_0$  siendo cierta) y  $1 - \beta / \Theta_1$  los errores de tipo II (rechazar  $H_A$  siendo cierta). Denotándose mediante  $\beta / A$  la restricción de la función  $\beta$  al conjunto  $A$ . Ante la imposibilidad de disminuir simultáneamente ambos errores una solución al problema es identificar pruebas que se caractericen por una tasa de error de tipo I específica y dentro de ellas tratar de minimizar el error de tipo II. Esto formalmente consiste en: Una  $\alpha$ -prueba es una prueba cuyas tasas de error de tipo I no rebasan  $\alpha$  ( $0 < \alpha < 1$ ), estas pruebas se definen a partir de la identificación de una región crítica,  $K_\alpha$ , para la cual la correspondiente función de potencia  $\beta(\theta) = P_\theta(\{\vec{X} \in K_\alpha\})$  ( $\theta \in \Theta$ ) cumpla  $\text{Sup}_{\theta \in \Theta_0} \beta(\theta) = \alpha$ . Este problema admite, en general, infinitas soluciones; es decir existen infinitas  $\alpha$ -pruebas o pruebas con un error de tipo I igual a  $\alpha$ ; dentro de la familia de las  $\alpha$ -pruebas se escoge aquella en la que el error de tipo II se minimice según algún criterio (17). De esta forma la Teoría de las Pruebas de Hipótesis de Neyman y Pearson introduce un orden o jerarquización entre los errores y por consiguiente entre las hipótesis.

Refiriéndose a los alcances del procedimiento de Pruebas de Hipótesis y sus repercusiones en el marco del análisis de los datos provenientes de la investigación, Neyman y Pearson señalaron “...ninguna prueba basada en la Teoría de Probabilidades puede, por sí misma, generar índices válidos sobre la verdad o la falsedad de una hipótesis. Las Pruebas de Hipótesis deben ser miradas con otra perspectiva. Siguiendo la regla de aceptar o rechazar una hipótesis no estamos diciendo nada definitivo sobre si la hipótesis es o no verdadera. Sin esperanza de saber cuándo una hipótesis por separado es cierta o es falsa debemos buscar reglas que gobiernen nuestro comportamiento respecto a ellas, siguiendo tales reglas podemos estar seguros que a la larga, como resultado de múltiples repeticiones de la experiencia (in the long run), no estaremos equivocados muy a menudo. ...” (18). Al ser las Pruebas de Hipótesis un procedimiento de decisión de naturaleza aleatoria y con probabilidades o tasas de error controladas, nunca pueden ser concluyentes de manera aislada. Desde posiciones metodológicas diferentes tanto Fisher como Neyman y Pearson coincidían en el papel, dentro de la metodología de la investigación, de sus procedimientos de análisis.

Dentro de la Teoría de las Pruebas de Hipótesis Neyman introdujo, en 1937, los Intervalos de Confianza como parte integrante de la Teoría General de las Pruebas de Hipótesis (6,19). Esta relación queda evidenciada en la definición de la noción, más general, de Conjunto de Confianza para estimar un parámetro  $\theta \in \Theta \subseteq \mathcal{R}^k$ . Para la definición se parte de un conjunto  $\Phi(\vec{X}) \subseteq \Theta$ , el que se denomina un conjunto de confianza para estimar  $\theta$  con un nivel de confianza  $1 - \alpha$  si  $P_\theta(\{\vec{X}: \theta \in \Phi(\vec{X})\}) = 1 - \alpha$ . En el caso en que  $k=1$ , o sea cuando el Espacio Paramétrico es unidimensional, un Conjunto de Confianza pudiera ser un intervalo con lo que el concepto de Intervalos de Confianza se determina como un caso particular de esta presentación teórica. La semejanza entre Pruebas de Hipótesis e Intervalos de Confianza se manifiesta en que, para la construcción de los Conjuntos de Confianza, se considera  $H_0: \theta = a$  para  $a \in \mathcal{R}^k$  fijo y se construye una  $\alpha$ -prueba para  $H_0$ , si  $K_\alpha$  representa la correspondiente región crítica, el conjunto  $\Phi(\vec{X}) = \{a \in \Theta: \vec{X} \in K_\alpha\}$  es un conjunto de confianza para estimar  $\theta$  con nivel de confianza  $1 - \alpha$  (20). En dependencia del tipo de alternativa considerada se obtienen entonces diferentes tipos de conjuntos de confianza.

Aunque Intervalos de Confianza y Pruebas de Hipótesis fueron concebidos como dos formas alternativas de dar respuestas a la misma pregunta no son exactamente equivalentes, metodológicamente son evidentemente diferentes. En ocasiones se emplean las Pruebas de Hipótesis sin identificar claramente los errores de tipo II, es decir sin

identificar claramente las características de la función de potencia, esa no es, ciertamente, la mejor manera de usarlas pues puede realizarse un análisis distorsionado. En este punto los Intervalos de Confianza pueden desempeñar un papel fundamental pues su amplitud, que representa los niveles de variabilidad de las estimaciones, da una medida de lo firme que pudiera resultar la aceptación o el rechazo de una hipótesis (21) al relacionarse directamente con los errores. La amplitud del Intervalo de Confianza es una medida de lo informativo que pudiera o no resultar el estudio y es el elemento más relevante en sus aplicaciones. Utilizar los Intervalos de Confianza para un proceso de decisión sin tener en cuenta su amplitud, es lo mismo que aplicar la regla de decisión que determina la correspondiente Prueba de Hipótesis desconociendo los errores. Emplear los Intervalos de Confianza como un instrumento alternativo a las Reglas de Decisión de Neyman y Pearson sin analizar la estimación del parámetro de interés que el intervalo presenta, es analizar los datos con un algoritmo divorciándose del problema específico y del significado particular de las mediciones que se analizan.

### La solución desde la estadística Bayesiana

La aplicación de criterios basados en el Teorema de Bayes al problema de validar Hipótesis Estadísticas, tuvo un considerable impulso en el año 1935 a partir de los desarrollos metodológicos de Harold Jeffrey para cuantificar la evidencia a favor o en contra de una Teoría Científica. En la solución Bayesiana al análisis de las Hipótesis Estadísticas se necesita, al igual que las Pruebas de Hipótesis de Neyman y Pearson, de una hipótesis alternativa aunque no se le da una función especial a ninguna de las dos hipótesis (22). Como parte fundamental del análisis Jeffrey introdujo el Factor de Bayes para pesar las evidencias a favor de una u otra hipótesis.

La solución Bayesiana se estructura sobre la misma base que los dos enfoques anteriores, a saber,  $\vec{\chi}$  representa el n-vector aleatorio de las observaciones,  $\Theta \subseteq \mathfrak{R}^k$  es el espacio paramétrico que determina la distribución de  $\vec{\chi}$ . Las Hipótesis Estadísticas de interés son  $H_0: \theta \in \Theta_0$ ,  $\Theta_0 \subset \Theta$  y  $H_A: \theta \in \Theta_1 = \Theta \setminus \Theta_0$ . La solución propuesta es, sin embargo, muy diferente a los dos enfoques anteriores pues, tanto Fisher, como Neyman y Pearson, se basan en una visión frecuentista de la Probabilidad; sin embargo, los procedimientos de la Estadística Bayesiana se obtienen a partir de la Probabilidad Subjetiva, es decir, en la visión Bayesiana de la Estadística se considera la probabilidad como una medida de la certidumbre de un observador sobre la plausibilidad de un suceso, siendo esta una medida del grado de convicción acerca de la plausibilidad o verosimilitud de un desenlace, en correspondencia con los conocimientos acumulados hasta el momento (23). La Estadística Bayesiana parte de

la aplicación del conocido Teorema de Bayes, denominado así en alusión a los resultados iniciales obtenidos por el Reverendo Thomas Bayes en el Siglo XVII para la obtención de probabilidades inversas; aunque fue Laplace en el siglo XIX quien desarrolló un sistema matemático de razonamiento inductivo, basado en la Teoría de Probabilidades, que en su esencia es reconocido como Bayesiano (24). Para el análisis de las Hipótesis Estadísticas la solución bayesiana consiste en suponer que el vector de observaciones  $\vec{\chi}$  se caracteriza por una ley de probabilidad  $\mu_{\vec{\chi}/\theta}$ , ley que es una medida imagen y que depende de  $\theta$ , valor desconocido que en este caso es considerado una variable aleatoria cuya ley se representa por la medida de probabilidad  $\mu$ . Esta distribución de probabilidad representa el conocimiento que se tiene a priori sobre  $\mu$  y se denomina distribución a priori o priors. Si no se dispone de información adicional se utilizan priors no informativos, es decir, priors que en general no jerarquicen los posibles valores de  $\theta$ . La ley condicional  $\mu_{\theta/\vec{\chi}}$  se denomina la distribución a posteriori del vector aleatorio  $\theta$  y representa el conocimiento sobre  $\theta$  después de realizado el estudio, se obtiene modificando la información a priori sobre el comportamiento de  $\theta$  con lo observado,  $\mu_{\vec{\chi}/\theta}$ . Esta inversión es posible gracias al Teorema de Bayes para vectores aleatorios, el que en su formulación más general establece que:

$$\frac{d\mu_{\theta/\vec{\chi}}}{d\mu} = \frac{P_{\vec{\chi}/\theta}}{\int P_{\vec{\chi}/\theta} d\mu} \quad [1]$$

Denotando  $\frac{d\mu_{\vec{\chi}/\theta}}{d\mu}$  la derivada de Radom-Nikodym y  $P_{\vec{\chi}/\theta}$  la función de densidad de la medida  $\mu_{\vec{\chi}/\theta}$ . En las aplicaciones más comunes, las leyes utilizadas en la formulación son medidas con densidad respecto a la medida de Lebesgue, lo que determina el caso continuo, o respecto a la medida Contadora, situación en la que queda determinado el caso discreto.

En pocas palabras la lógica de este procedimiento radica en que las distribuciones de probabilidad representan el conocimiento sobre los parámetros considerados en las Hipótesis Estadísticas. Partiendo de un conocimiento inicial o a priori sobre los posibles valores de esos parámetros, los priors, este conocimiento es modificado por los datos y esta modificación es representada por la distribución a posteriori. La posibilidad de utilizar los conocimientos previos en el análisis es una de las fortalezas de la Estadística Bayesiana, pero es a la vez una de sus debilidades al no existir una fuente muy clara acerca de cómo construir la distribución a priori; existiendo el peligro de obtener conclusiones sesgadas. Una situación extrema en la selección de los priors se presenta en la paradoja de Lindley en la que  $H_0$  siempre es aceptada (25), aunque la relevancia práctica de tal paradoja ha sido cuestionada por algunos autores (26).

Para el proceso de valoración de las dos hipótesis Jeffrey propuso utilizar el factor de Bayes (27)

$$B(H_0, H_A) = \frac{P(\{\theta \in \Theta_0 / \vec{\chi}\})}{P(\{\theta \in \Theta_1 / \vec{\chi}\})} \cdot \frac{\mu_\theta(\Theta_0)}{\mu_\theta(\Theta_1)} \quad [2]$$

El que, de manera más simplificada, puede escribirse como:

$$B(H_0, H_A) = \frac{\mu_\theta(\Theta_1) \int_{\Theta_0} P_{\vec{\chi}} / d\mu_\theta}{\mu_\theta(\Theta_0) \int_{\Theta_1} P_{\vec{\chi}} / d\mu_\theta} \quad [3]$$

El Factor de Bayes representa el cociente de dos razones. El numerador es la razón de las probabilidades a posteriori de las hipótesis  $H_0: \theta \in \Theta_0$  y  $H_A: \theta \in \Theta_1$ , el denominador la correspondiente razón pero con las probabilidades a priori. Para su interpretación, Jeffrey introdujo un criterio que gradúa las evidencias a favor de  $H_0$  o de  $H_A$  en fuertes, moderadas y débiles en correspondencia con los valores de  $B(H_0, H_A)$  (22).

Como parte del Análisis Bayesiano de las Hipótesis Estadísticas se utilizan también los Intervalos de Credibilidad. Estos intervalos permiten estimar los parámetros de interés bajo el paradigma Bayesiano, por lo que de nuevo en este caso son considerados variables o vectores aleatorios. Los Intervalos de Credibilidad se oponen a los Intervalos de Confianza frecuentistas de los que constituyen una alternativa más racional: Mientras un intervalo de confianza contiene al parámetro como un atributo determinado por múltiples repeticiones del experimento, noción extremadamente difícil de entender para muchos, el Intervalo de Credibilidad se construye con el criterio de que cada intervalo obtenido contiene al valor de interés con una probabilidad subjetiva a posteriori dada.

De manera general una Región de Confianza p-creíble  $C_p(\vec{\chi})$  ( $0 \leq p \leq 1$ ) es cualquier subconjunto de  $\Theta$  tal que

$$\int_{C_p(\vec{\chi})} d\mu_{\theta/\vec{\chi}} = p \quad [4]$$

El Intervalo de Credibilidad es un caso particular de las Regiones de Confianza p-creíbles cuando  $\Theta \subseteq \mathfrak{R}$ .

A pesar de la coherente de su definición e interpretación, la existencia de infinitas regiones p-creíbles pudiera dificultar su uso al introducir algunas inconsistencias en las estimaciones, sobre todo en el caso en que las distribuciones a posteriori fueran marcadamente asimétricas. Por su facilidad de interpretación y la coherencia del criterio para su construcción frente a las asimetrías de la distribución, las regiones de máxima densidad son muy utilizadas.

## Las críticas a las pruebas de hipótesis y las pruebas de significación. Sus raíces históricas y metodológicas

La situación actual es que múltiples publicaciones científicas desestimulan o prohíben las Pruebas de Hipótesis, creándose un contrasentido pues, una mirada a cualquier programa de un curso de Estadística, permitirá afirmar que esos contenidos son los de mayor presencia. ¿Cómo es posible que uno de los temas más tratados en los cursos de Estadística y con mayor presencia en muchos libros de texto estén proscriptos para ser usados en las publicaciones científicas? La respuesta a esta pregunta es compleja y tiene múltiples aristas.

Una primera componente en la explicación del problema se obtiene al dar una mirada crítica a la forma en que se aplican las Pruebas de Hipótesis. La práctica usual consiste en comparar el valor P -una medida de evidencia- con el valor  $\alpha$  -una tasa de error- que en la inmensa mayoría de los casos es fijada como 0,05; el criterio se basa en utilizar un algoritmo de conducta en el que si  $P < \alpha$ , comúnmente  $\alpha = 0,05$  se decide rechazar  $H_0$  y en el caso contrario se considera válida la alternativa. Esta forma de proceder mezcla, de manera ecléctica, los enfoques antagónicos de Fisher y de Neyman-Pearson. Es cierto que, desde el punto de vista algebraico, comparar P con  $\alpha$  es equivalente a considerar la región crítica de Neyman y Pearson en el proceso de decisión pues, si  $K_\alpha$  es la región de rechazo de una  $\alpha$ -prueba el valor P, determinado por un vector de observaciones  $\vec{\chi}$ , puede obtenerse por  $P(\vec{\chi}) = \inf \{\alpha: \vec{\chi} \in K_\alpha\}$ , esto significa que el valor P es el menor valor del error de tipo I,  $\alpha$ , para el cual la hipótesis nula pudiera ser rechazada. El problema está en que lo que se hace en el análisis de los datos de la investigación no es álgebra solamente y por ello esta forma de ver el análisis de las hipótesis es errónea: no es correcto comparar una medida de evidencia con una tasa de error, son dos números metodológicamente diferentes, más aun incompatibles al proceder de los criterios de análisis de Fisher y de Neyman-Pearson, modelos diferentes e históricamente antagónicos (6,13, 29). Por encima de la concordancia algebraica hay un problema con diferentes aristas, al aplicar esa regla de decisión, por una parte se están comparando dos números con significaciones metodológicas diferentes y por otra el interés se centra en la propia desigualdad, no en el valor P y lo que él representa (6). No hay una idea clara acerca de cómo comenzó a suceder esto, ni es posible identificar claramente un momento inicial para esta práctica que rige la conducta tanto de los profesionales en estadística como de los investigadores de otras ramas del conocimiento que la utilizan. Lamentablemente se ha extendido, de hecho muchos libros de texto toman posiciones inespecíficas sobre este tema, profundizando así en los problemas al ser



aplicadas las Pruebas de Hipótesis. El proceso de mezclar los enfoques de Neyman-Pearson y de Fisher surgió con el transcurso del tiempo e irónicamente ocurrió a pesar de las diferencias metodológicas de ambos enfoques y por encima de las diferencias metodológicas y personales irreconciliables entre Neyman y Fisher.

Resulta sorprendente el ritual de considerar en casi todos los problemas el valor 0,05. Parece un problema religioso más que científico el que una misma tasa de error sea válida para cualquier situación. Desde sus puntos de vista antagónicos tanto Fisher como Neyman y Pearson resaltaron la necesidad de no hacer esto y ser flexibles en los análisis. Es trágicamente gracioso el que los populares valores de 0,05 y el menos conocido 0,01 para las inferencias fueron establecidos por Fisher en su libro *Statistical Methods for Research Workers* -Sir Ronald A. Fisher. Edinburgh (Oliver and Boyd), 12th Ed., 1954. Pp. xv, 356; 12 Figs., 74 Tables. 16s- básicamente porque calculó sus tablas con esos valores por las restricciones editoriales y de derecho de autor de una publicación científica.

La aplicación algorítmica de las Pruebas de Hipótesis se ha convertido en un ritual bastante pernicioso. Según refirió Yates en el lejano 1951 "... esto ha causado que los investigadores presten una atención indebida a los resultados de las Pruebas de Significación que realizan con sus datos y muy poca a la estimación de la magnitud de los efectos que están estimando..." (30). Es bastante frecuente que el proceso de análisis se reduzca a aplicar una Prueba de Hipótesis y marcar, en cada variable analizada, si  $p < 0,05$  o no, clasificando, a partir de ahí, los resultados en "significativos o no". En muchos casos sobre esta base se pasa a análisis más complejos, no se hace nada más para comprender el comportamiento de las variables; perdiendo así el investigador la posibilidad de comprender a profundidad el significado de sus hallazgos. Algunas personas tratan de evadir los problemas de las Pruebas de Hipótesis sustituyéndolas por intervalos de confianza pero al final construyen la misma regla de decisión según el parámetro esté o no en el intervalo.

Los libros de texto presentan las Pruebas de Hipótesis como una disciplina bien estructurada, como un conocimiento bien establecido e inamovible (13); en ellos las Pruebas de Hipótesis son presentadas con objetividad y exactitud. Bajo el amparo de las correspondientes formulaciones matemáticas, esto hace creer a muchos especialistas de otras áreas que usarlas los protege con la exactitud de las Ciencias Exactas y que por lo tanto sus procesos de análisis son infalibles. La no comprensión exacta de los procedimientos que se aplican "...da un aire de objetividad

científica..." como fue señalado por Carver en 1978 (31). Bajo este manto matemático "formal" queda oculto el que muchas de las pruebas en uso tienen propiedades optimales en una clase de pruebas muy reducida y el que, en otros casos, las pruebas utilizadas sólo tienen un fundamento intuitivo (20). Estos sesgos en los razonamientos inferenciales y el exceso de confianzas derivado del sustrato matemático de los procedimientos podrían explicarse, como señaló Batanero, "...por el pobre razonamiento de los adultos en problemas probabilísticos..." (32).

En múltiples ocasiones los investigadores tampoco tienen un conocimiento preciso de qué significa el valor P con el que toman sus conclusiones, en muchos casos piensan que es la probabilidad de que la hipótesis nula sea verdadera (12). Lo que es un reflejo de que, en el aspecto docente, el problema con las Pruebas de Hipótesis es un problema compartido entre estudiantes y profesores (33).

Los procedimientos actuales para enseñar la estadística a profesionales de otras áreas y los propios libros de texto han contribuido a la situación planteada. Muchos de los libros de texto utilizados para esos fines son reediciones de libros de años anteriores y como tal fueron escritos en otro contexto en el que las Pruebas de Hipótesis no habían mostrado sus inconsistencias metodológicas (34). En general estos libros presentan los algoritmos de conducta para enfrentar los problemas de Pruebas de Hipótesis, promoviendo de esta forma los rituales del algoritmo para aceptar o rechazar  $H_0$  mezclando de una u otra forma los enfoques de Neyman-Pearson con el de Fisher. No promueven un análisis integral de los datos, con flexibilidad y a la medida del problema, combinando diferentes criterios de análisis que permitan, en su conjunto, una comprensión total del problema. Esto quizás es un remanente de los mecanismos de enseñanza de las matemáticas a no matemáticos, donde la función principal es repetir hasta la saciedad problemas idénticos, clasificados en secciones temáticas independientes, buscando una habilidad para resolverlos con la que, paradójicamente, se genera un esquematismo que aleja a las personas de las posibilidades de aplicación. En ese sentido se impone un cambio en los procedimientos pedagógicos tanto de la Estadística como en las Matemáticas, donde la función principal debe ser acercar a los estudiantes a los problemas y sus soluciones. Dentro de esta estrategia de perfeccionamiento de la Enseñanza de las Matemáticas, la Estadística pudiera ser una parte importante al mostrar cómo modelar problemas donde la incertidumbre juega un papel relevante (34). En el caso de la formación de los profesionales en Estadística la formalización matemática en la presentación de los contenidos, unido a un divorcio de los aspectos prácticos, pudiera ser un elemento central en la explicación del problema.

Otro elemento que permitiría explicar los problemas con las Pruebas de Hipótesis desde la componente docente es que la lógica de las Pruebas de Hipótesis y su función metodológica en la investigación son difíciles de entender, sobre todo para estudiantes de los primeros años de las carreras, momento en el que se dictan las asignaturas de estadística y en el que ellos no están familiarizados con la investigación (35).

En la aplicación de las Pruebas de Hipótesis existe una arista metodológica que resulta extremadamente relevante. Una mirada a las publicaciones científicas donde estas son aplicadas mostrará que existen un grupo de pruebas –paramétricas, asociadas a la distribución normal- que se presentan profusamente en las aplicaciones. Esta estandarización refleja, en otro aspecto del análisis de los datos, el mismo esquematismo que se presenta en el algoritmo para la aplicación de las Pruebas de Hipótesis. En este sentido coexisten una limitada creatividad en la búsqueda de los procedimientos de análisis con deficiencias en la identificación de las hipótesis y en la determinación de su significado en el contexto del problema a analizar. Por ejemplo, según la experiencia personal del autor en un problema donde se comparan dos poblaciones normales, la comparación de varianzas es utilizada en muchos casos sólo como un punto de inflexión para decidir qué prueba de comparación de medias utilizar; en ese contexto y sin ningún rubor, el investigador concluye que los grupos son iguales porque las medias lo son a pesar de que las varianzas sean diferentes. Otro ejemplo, también tomado de la experiencia del autor, entre los múltiples que se podrían citar de esquematismos en los análisis, es el que muchos investigadores buscan relevancia en la relación entre dos variables continuas aplicando la Prueba de Hipótesis para analizar la hipótesis nula de que el coeficiente de correlación es 0, interpretándose el que se rechace esta hipótesis como sinónimo de haber encontrado una relación relevante, pero eso no necesariamente es así, en ese caso sólo se encontró una relación no nula; la que pudiera no ser todo lo relevante que el investigador estaba buscando y en consecuencia ser espuria a pesar de “su significación”. Los aspectos de la identificación de las hipótesis relevantes al problema, comprender lo que realmente significa el aceptarlas o rechazarlas, así como la selección de la prueba estadística más adecuada a lo que se desea analizar son puntos muy relevantes dentro del diseño del análisis estadístico de la investigación y son los posibles causantes de algunas de las inconsistencias que se presentan en el análisis de los datos.

Muchas de las hipótesis utilizadas comúnmente buscan la igualdad, ya sea a 0 o entre grupos, este tipo de

hipótesis fue denominado por Cohen en 1994 “hipótesis nada” -null hypothesis- (30) existe un consenso en que estas hipótesis son poco relevantes en un problema de investigación, siendo este consenso bien claro en el campo de los Ensayos Clínicos. Para entender la irrelevancia de estas hipótesis bastaría referirse a una afirmación de Tuckey citada por Cohen (30) “...es tonto preguntar si los efectos de A y B son diferentes, ellos siempre lo son en algún lugar decimal...”. Esto se refleja en que, al ser dos números diferentes, cuando el tamaño de muestra aumenta el error de muestreo de las estimaciones disminuye y en consecuencia también el efecto del azar por lo que cualquier prueba de hipótesis va a detectar diferencias a partir de cierto tamaño de muestra. Las hipótesis de tipo puntual se sabe que son falsas al inicio de la investigación, en realidad si no se rechazan es porque el tamaño de muestra es pequeño (36, 37); tal vez esta sea su función, el control primario del efecto del azar, pero este aspecto se torna metodológicamente difícil de entender y manejar por los investigadores.

En general en la formulación de las hipótesis no se parte de un conocimiento profundo de las mediciones en estudio y del significado de sus valores, por ello no se identifican umbrales a partir de los cuales los valores de las variables o sus diferencias resulten relevantes en el problema analizado. En general se escogen las “hipótesis nada” correspondientes y se procede con el software de análisis disponibles a “analizar los datos”. En muy pocos casos ellas son sustituidas por hipótesis de equivalencia, superioridad o inferioridad con lo que la comparación de las poblaciones y la identificación de la relevancia de las mediciones tendría un sentido práctico y en correspondencia las conclusiones serían más realistas. Por ejemplo, al comparar dos poblaciones el problema no es preguntarse si son diferentes sino si las diferencias existentes son relevantes para las medición en estudio; así, si usted compara el peso de dos poblaciones de humanos adultos, el problema a analizar no es determinar si son diferentes sino establecer si la magnitud de esas diferencias es digna de ser considerada desde el punto de vista nutricional para concluir que existe una diferencia en el estado nutricional de los dos grupos: una diferencia, por ejemplo, de medio kilogramo es irrelevante aunque sea “estadísticamente significativa”. Esta forma de proceder abriría las puertas a análisis de datos a la medida de cada problema y no patrones genéricos de análisis que se utilizan en cualquier circunstancia. La determinación de tales valores de umbral es un aspecto que aguarda por la atención de los investigadores de las diferentes áreas que deben enfrentar el reto de conocer más a profundidad el significado de sus mediciones, tal y como fue señalado por Feinstein en 1996 (38).

## Los retos docentes y metodológicos para la aplicación de las Pruebas de Hipótesis

Existe una tendencia a desechar las Pruebas de Hipótesis, esta se evidencia en las restricciones editoriales mencionadas anteriormente. Ciertamente no parece conveniente renunciar a su objetividad porque, como señaló Fleiss, "...ellas tienen la virtud de ser criterios explícitos y pre-especificados para las inferencias..." (39); es cierto que han sido muy mal utilizadas y se ha abusado de ellas, pero "...la no interpretación correcta, su mal uso, no es un argumento para prohibirlas..." (40). En este sentido se impone sustituir el ritual de las Pruebas de Hipótesis por procesos de análisis de los que ellas formen parte, estos procesos deben ser estructurados a partir de la formulación de las hipótesis a la medida del problema, considerando las características de las mediciones y debe promover un análisis integral de los datos tomando como punto de partida los hallazgos de la fase exploratoria. El cumplimiento de esta aspiración partiría de lograr una mayor vinculación entre el planteamiento metodológico de la investigación y el análisis estadístico, entendido este como el muestreo y el análisis de los datos.

Es importante seguir la visión de Fisher y su valor P en la investigación científica, esto significa considerar las Pruebas de Hipótesis como parte de un análisis integral en el que se deben combinar las evidencias obtenidas en los datos con información adicional, externa al estudio (41). El enfoque de Neyman y Pearson pudiera ser muy relevante en problemas de Control Estadístico de la Calidad, pero en el caso de la investigación ha quedado demostrado que metodológicamente no es adecuado (12).

En las publicaciones en que se restringe el uso de las Pruebas de Hipótesis se promueve el uso de los Intervalos de Confianza, es cierto que ambos deben ser utilizados como partes de un análisis integral pues brindan informaciones complementarias, pero es importante no sustituir el ritual de las Pruebas de Hipótesis por el ritual de construir una regla de decisión utilizando el Intervalo de Confianza, de ser así nada aportarían estos últimos. En lugar de buscar un nuevo ritual mágico se debe profundizar en la fase exploratoria de los datos, promoviendo el uso de las técnicas del Análisis de Datos Exploratorio y complementar la aplicación de las Pruebas de Hipótesis con el análisis de la potencia (30), para ello pueden utilizarse, de manera complementaria, los Intervalos de Confianza; los que deben ser analizados en ese contexto y teniendo en cuenta la magnitud de la estimación que brindan, su significado en el contexto de la medición a analizar, así como el efecto del azar sobre esas estimaciones. En la presentación de los resultados de la investigación, específicamente en las publicaciones científicas, se debe evitar presentar el resultado de las pruebas estadísticas con

juicios que resumen la "significación" de las diferencias como NS, \*, \*\*,  $p < 0,05$ ; etc; los que ha quedado demostrado son inadecuados. Lo relevante debe ser presentar explícitamente la medida de la evidencia contra  $H_0$  obtenida en el estudio, el valor P y analizarlo siguiendo el criterio de interpretación presentado por Sterne y Smith (12) quienes consideraron un gradiente en la evidencia aportada por esos valores tomando como base la propuesta de interpretación del valor P de Burdette y Geham que aparece referida en Royal (37). El análisis del valor P con este criterio debe realizarse tomando como marco de referencia el tamaño de la muestra (37) y teniendo en cuenta lo señalado por Selwyn en 1989 "...los valores P son objetivos pero a menos que tengan valores muy extremos ellos deben ser interpretados a la luz de otra información..." (42)

Como parte del cambio en la mentalidad para hacer el análisis de datos está el promover entre editores de revistas, asesores de tesis y oponentes de las mismas la necesidad de desprenderse de las Pruebas de Hipótesis como elemento final en el análisis de los datos y entender este como un conjunto de acciones que no necesariamente deben conducir a la realización de las mismas (8). En el caso en que estas sean aplicadas se debe proceder según la naturaleza aleatoria y el significado del valor P.

En lo docente el reto a asumir es grande, es necesario promover la elaboración de textos que permitan adiestrar para la ejecución de estos criterios para el análisis de los datos. En ellos debe hacerse énfasis en el uso adecuado de los Intervalos de Confianza, resaltando lo que realmente ellos pueden añadir al análisis de los datos, asimismo deben diferenciar los enfoques de Neyman - Pearson del de Fisher, evitando las amalgamas que han resultado ser tan perniciosas. Los programas docentes deben promover el uso de simulaciones para profundizar en el conocimiento de los estudiantes en temas como variabilidad muestral y distribuciones muestrales (33) mejorando así la baja percepción que se logra en la actualidad de lo que es el azar y sus efectos. En el caso de los valores P las simulaciones pudieran ayudar a comprender en realidad su esencia aleatoria, en ese sentido pueden perfeccionarse los criterios de enseñanza a partir de las recomendaciones de Murdoch y colaboradores (43). La incorporación en los cursos de estadística del tratamiento de las hipótesis de equivalencia, superioridad y no inferioridad será un punto central para promover análisis de datos a la medida de los problemas y sus variables; para ello es necesario promover estudios que identifiquen los valores de umbral para las diferentes mediciones en el marco de su significado y de los errores de medición que se cometen en su obtención pues, sucede incluso, que en muchos casos mediciones que se obtienen con error son tratadas como si no lo fueran. La profundización en el estudio de los errores

de medición, su efecto sobre los procedimientos de análisis de los datos y cómo incorporarlos al análisis, debe mejorar la enseñanza de la estadística y tendrán un efecto inmediato sobre la calidad de la investigación al permitir aumentar la potencia en la aplicación de las pruebas de hipótesis. Según Sterne y Smith la disminución de los errores de medición es una forma de aumentar la potencia sin necesidad de aumentar el tamaño de muestra (12).

Partiendo de la relatividad de la evidencia presentada por cada estudio es importante promover el uso de los metaanálisis para integrar los resultados de diferentes estudios, para ello deben promoverse investigaciones con Diseños Metodológicos comparables para que la integración de estudios en la búsqueda de respuestas a los problemas de investigación sea más fructífera

## Conclusiones

A pesar de las críticas a su uso las Pruebas de Significación siguen siendo útiles en el análisis de los datos de la investigación pero es importante promover que sean utilizadas de manera correcta en correspondencia con sus alcances y funciones metodológicas. Se necesita determinar las Hipótesis Estadísticas y las Pruebas para validarlas según las características de las mediciones y el problema a analizar. En ese sentido es importante promover el uso de las hipótesis de superioridad, equivalencia y no inferioridad. Se deben introducir modificaciones en la Enseñanza de la Estadística para capacitar a quienes la aplican para realizar un análisis integral de los datos, en el que las Pruebas de Hipótesis o de Significación constituyan una parte integrante del mismo pero no el todo; evitando los rituales generados por el uso de algoritmos para decidir a qué conclusiones arribar en el análisis e identificando los alcances metodológicos y las limitaciones de las conclusiones a que se arribe.

## Financiación

Este artículo fue financiado por la Facultad de Ciencias de la Universidad del Rosario.

## Conflicto de intereses

El autor no presenta conflicto de intereses con este trabajo.

## Referencias

1. American Statistical Association. Careers in Statistics. What is Statistics. <http://www.amstat.org/careers/whatisstatistics.cfm>. Consultado 9 de julio de 2012.
2. Polit D, Hungler P. Investigación Científica en Ciencias de la Salud. Principios y Métodos. Sexta Edición. McGraw-Hill Interamericana, Health Care Group, México DF. 2000, 715p
3. International Comitee of Medical Journal Editors. Uniform requirements for manuscript submitted to biomedical journals. *Br Med J*. 1997; **336**(4):309-15.
4. Epidemiology. Instructions for Authors. <http://edmgr.ovid.com/epid/accounts/ifauth.htm>. Consultado 9 de julio de 2012
5. DeGroot M. A Conversation with CR Rao. *Statistical Science* 1987; **2**(1):53-67
6. Moran JL. A Farewell to P-value? *Critical Care and Resuscitation* 2004; **6**:130-137
7. Gigerenzer G. We Need Statistical Thinking, not Statistical Rituals. *Behavioral and Brain Sciences* 1998; **21**(2):194-195
8. Chatfield C. The Initial Examination of Data. *Journal of the Royal Statistical Society. Series A (General)* 1985; **148**(3): 214-253
9. Perrin E. On the Importance of Exploring Data (Editorial). *Medical Care* 1981; **XIX**(12):1163-1164
10. Rozenkranz G. Hypothesis. *Encyclopedia of Clinical Trials*. John Wiley and Sons, 2007
11. Rozeboom W. The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin* 1960; **57**(5): 416-428
12. Sterne JA, Smith GA. Sifting the Evidence-What's Wrong With Significance Tests? *British Medical Journal* 2001; **322**: 223-231
13. Hubbard R y Bayarri MJ. Confusion Over Measures of Evidence ( $p$ 's) Versus Error ( $\alpha$ 's) in Classical Statistical Testing. *The American Statistician* 2003; **57**(3):171-182
14. Lambert D, Hall WJ. Asymptotic Lognormality of P-values. *The Annals of Statistics* 1982; **10**(1):44-64
15. Hung HM, O'Neill R, Bauer P, Köhne K. The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics* 1997; **53**:11-22
16. Sackrowitz H, Samuel-Cahn E. P Values as Random Variables-Expected P Values. *The American Statistician* 1999; **53**(4): 326-331
17. Lehmann EL. Testing Statistical Hypothesis. John Wiley and Sons, 1959, p369
18. NeymanJ, Pearson E. On the Problem of the Most Efficient Test of Statistical Hypothesis. *Trans of the Royal Society of London* 1933;A231:289-337

19. Neyman J. Frequentist Probability and Frequentist Statistics. *Synthese* 1937; **36**:97-131
20. Rao CR. *Linear Statistical Inference and Its Applications*. John Wiley and Sons 1973, p625
21. Natrella M. The Relation Between Confidence Intervals and Test of Significance –A Teaching Aid. *The American Statistician* 1960; **14**(1):20-38
22. Kass R, Raftery A. Bayes Factor. *Journal of the American Statistical Association* 1995; **90**(430):773-95
23. Lee P. *Bayesian Statistics An Introduction*. Third Edition. Hodder Arnold, London 2003, 351p
24. Dale A. Bayes or Laplace? an examination of the origin and early application of Bayes' theorem. *Archive for the History of the Exact Sciences* 1982; **27**: 23–47.
25. Christiansen R. testing Fisher, Neyman, Pearson and Bayes. *The American Statistician* 2005; **59**(2):121-126
26. Berger J, Delampady M. Testing Precise Hypthesis. *Statistical Science* 1987; **2**(3): 317-352
27. Robert C. Le Choix Bayésien. *Principles et Pratique*. Springer, Paris 2006, 638p
28. Bernardo J. Un Programa de Síntesis para la Enseñanza Universitaria de la Estadística Matemática Contemporánea. *Revista de la Real Academia de Ciencias Exactas Físicas y Naturales* (España) 2002; **95**(1-2):81-99
29. Lenhard L. Models and Statistical Inference: The controversy between Fisher and Neyman-Pearson. *The British Journal for the Philosophy of Science* 2006; **57**: 69-80
30. Cohen J. The Earth is Round ( $p < .05$ ). *American Psychologist* 1994; **49**(12):997-1003
31. Carver R. The Case Against Statistical Significance Testing Revisited. *Journal of Experimental Education* 1978; **61**(4): 287-292
32. Batanero C. Controversies Around the Role of Statistical Tests in Experimental Research. *Mathematical Thinking and Learning* 2000; **2**(1&2):75-97
33. Sohlberg S, Andersson G. Extracting a Maximum of Useful Information from Statistical Research Data. *Scandinavian Journal of Psychology* 2005; **46**:67-77
34. Greer B. Statistical Thinking and Learning. *Mathematical Thinking and Learning* 2000; **2**(1&2): 1-9
35. Gliner J, Leech N, Morgan G. Problems With Null Hypothesis Significance Testing (NHST): What do Textbooks Say?. *The Journal of Experimental Education* 2002; **71**(1):83-92
36. Johnson D. The Insignificance of Statistical Significance Testing. *Journal of Wild Life Management* 1999; **63**(3):763-772
37. Royal R. The Effect of Sample Size on the Meaning of Significance Test. *The American Statistician* 1986; **40**(4): 313-314
38. Feinstein A. Two Centuries of Conflict-Collaboration Between Medicine and Mathematics. *Journal of Clinical Epidemiology* 1996; **49**(12):1339-1343
39. Fleiss J. Significance Tests Have a Role in Epidemiologic Research: Reactions to A.M. Walker. *American Journal of Public Health* 1986; **76**(5): 559-560
40. Weinberg C. It's Time to Rehabilitate P-Value. *Epidemiology* 2001; **12**(3): 288-290
41. Goodman S. Toward Evidence-Based Medical Statistics. 1. The P Value Fallacy. *Annals of Internal Medicine* 1999; **130**: 995-1004
42. Selwyn M. Dual Controls, p-Value Plots, and the Multiple Testing Issue in Carcinogenesis Studies. *Environmental Health Perspectives* 1989; **82**:337-344
43. Murdoch D, Tsai YL, Adcock. P-Values are Random Variables. *The American Statistician* 2008; **62**(3):242-245