Procedimientos de Comparaciones Múltiples

Edgar Steven Baquero Acevedo

December 1, 2020

1 Introducción

Parte del estudio dela inferencia estadística está relacionada con el cuestionamiento del espacio de parámetros que estimamos a través de los datos. Es así, como la formalización de una pregunta se vuelve conveniente desde el punto de vista teórico, pues nos permite abordar preguntas con el rigor necesario para tomar una desición. Esta formalización puede ser llevada a cabo por procedimientos estadísticos con el fin de juzgar si una propiedad se satisface para una población, con base en lo observado por una muestra de dicha población.

El procedimiento anteriormente nombrado, es conocido como *prueba de hipótesis*, y mediante esta teoría, es posible abordar problemas estadísticos, considerando una hipótesis nula y una alternativa, que luego definiremos con más detalle. Extendiendo un poco más este concepto, nos encontramos con la teoría de pruebas de hipótesis múltiples (PHM o MCP por sus siglas en inglés), de la cual, se ha desarrollado mucha investigación, sobretodo en los últimos años, donde ha presentado su auge en trabajos relacionados al de Benjamini y Hochberg (Ver ——).

El procedimiento se presenta cuando consideramos un conjunto de inferencias de manera simultánea (Ver por ejemplo —-)o se infiere un subconjunto de parámetros basados en valores observados (Ver por ejemplo ——);procedimiento que presenta algunos inconvenientes entre más inferencias son hechas, ya que es más probable realizar una inferencia errónea. Sin embargo, técnicas han sido desarrolladas con el fin de prevenir esto, permitiendo comparar niveles de significancia para una y más pruebas de manera directa. Estas técnicas generalmente requieren de un umbral de significancia más estricto para pruebas individuales, a cambio de poder compensar el umbral general de una prueba múltiple.

2 Formalización

2.1 Pruebas de hipótesis simples

Definimos una prueba de hipótesis *simple* como una prueba en la que interviene sólo una única hipótesis nula H_0 y su complemento, la hipótesis alternativa H_1. Se denomina *simple* puesto que sólo se tiene una conjetura a probar. El caso

donde se abordan más de una conjetura será objeto de estudio de la siguiente sección.

El ejemplo más común en la literatura a una prueba de hipótesis simple, es un juicio oral en el cual un ciudadano es acusado de un crimen particular. En dicha situación, el fiscal tratará de probar la culpabilidad del acusado y, sólo cuando haya suficiente evidencia para ello, éste será condenado. El juez, en este caso, se enfrentará a un problema donde intervienen dos hipótesis: H_0 : El acusado es inocente y H_1 : El acusado es culpable. Nótese la importancia conceptual del orden de la elección de las hipótesis, la hipótesis nula es siempre la hipótesis que se encuentra en prueba directa y cuya veracidad no se está dispuesto a rechazar a menos que haya evidencia suficiente para ello. En el caso del juicio, el acusado permanecerá siendo inocente a menos que haya evidencia suficiente para asumir lo contrario. Visto de esta forma, el juez no quisiera rechazar la hipótesis nula a menos que haya evidencia contundente para ello; rechazarla cuando en realidad es cierta constituiría un error grave pues se enviaría a un individuo inocente a prisión. En el contexto de pruebas de hipótesis este error se conoce como *error tipo I* y es de especial importancia controlar las posibilidades de que ocurra. Análogamente, si el juez decide que no existe evidencia suficiente para condenar al acusado siendo que éste en realidad es culpable estaría cometiendo otro tipo de error, quizás subjetivamente de menor impacto que el error tipo I, que en el contexto de pruebas de hipótesis se denomina *error tipo II*. Generalmente, la elección del orden de H_0 y H_1 se fija de acuerdo con el contexto y se hace de tal manera que reducir el error tipo I sea de mayor prioridad que reducir el error tipo II.

	H_0 Cierta (inocencia)	H_0 Falsa (Culpable)
H_0 rechazada (Condenado)	Error tipo I	Decisión correcta
H_0 no rechazada (Libre)	Decisión correcta	Error tipo II

A pesar de que el anterior ejemplo nos presenta de manera natural el surgimiento del tipo de errores, nos induce de manera intuitiva (y erróneamente) que el error tipo I y tipo II no están relacionados, pero es posible demostrar que reducir de manera simultánea ambos tipos de error no será posible pues reducir uno de ellos aumentará el otro, como se especifica a continuación:

$$P(\textit{error tipo I}) \to 0 \quad \Longrightarrow \quad P(\textit{error tipo II}) \to 1$$
$$P(\textit{error tipo II}) \to 0 \quad \Longrightarrow \quad P(\textit{error tipo I}) \to 1$$

En la mayoría de problemas donde se trabaja con pruebas de hipótesis, nos interesamos en disminuir el error tipo I ya que se presenta con mayor prioridad generalmente; razón por la cual, es conveniente controlarlo. Dado que en la práctica, llevar este error a 0 resulta poco práctico, establecemos una cota superior para la cual este puede ser encontrado. Dicha cota se conoce como *nivel de significancia* y se denota con la letra α. Una vez se asegura que un procedimiento de prueba de hipótesis cumple con un nivel de significancia fijo, es de

interés controlar el error tipo II y el proceso de control de este error son conocidas como *Pruebas Uniformemente Potentes*, cuyos detaller técnicos pueden ser revisados en Casella and Berger (2008).

Una vez establecida la lógica natural de una prueba de hipótesis simple, procedemos a formalizar algunos de los términos que la componen.

Hipótesis de trabajo: La aseveración acerca del espacio parametral Θ que nos interesa probar, junto con su complemento o hipótesis alternativa. Usualmente es denotado de la siguiente forma:

$$H_0 : \theta \in \Theta_0; \quad H_1 : \theta \in \Theta_1,$$

donde $\{\Theta_0, \Theta_1\}$ es una partición de Θ, el espacio de parámetros.

Estadístico de prueba: Valor calculado en función de la muestra observada, frecuentemente para resumir la información contenida en los elementos observados para propósitos comparativos. La elección del estadıístico de prueba conveniente es crucial en toda prueba de hipótesis y por lo general se hace con base en el contexto. El estadıístico de prueba escogido debe ser aquel que recoja información de la muestra y sea capaz de dar evidencia, mediante una distribución de probabilidad definida, en contra de la hipótesis nula que pueda ser cuantificada. Generamente lo denotamos por T.

Región de Rechazo (C): Región del espacio muestral, i.e. el conjunto de valores que puede tomar el estadıístico de prueba T , para los cuales se rechaza la hipótesis nula. En otras palabras, H_0 será rechazada si y sólo si, ocurre el evento $\{T \in C\}$. La forma de C puede depender, entre otras cosas, del tamaño muestral, de la distribución de T y del tipo de prueba de hipótesis que se está llevando a cabo.

Potencia: Es la medida de la capacidad de una prueba particular de rechazar correctamente la hipótesis nula. Es decir, la probabilidad de no cometer error de tipo II. Frecuentemente la potencia de una prueba se suele describir en términos de la llama función potencia que se define como la probabilidad de rechazar la hipótesis nula en función del valor verdadero del parámetro:

$$\beta(\theta^*) = P(T \in C | \theta = \theta^*) \tag{1}$$

Visto de esta manera, esperarıíamos que, para que una prueba sea de buena calidad experimental, $\beta(\theta)$ sea lo más pequeña posible para los valores $\theta \in \Theta_0$ y cercana a uno para los valores $\theta \in \Theta_1$.

Nivel de significancia(α): Como lo comentamos antes, es el mayor valor posible para la probabilidad de error de tipo I que el investigador está dispuesto a tolerar. En términos de la función de potencia definida en (1), se define mediante la siguiente expresión:

$$\alpha \leq \sup_{\theta \in \Theta_0} \beta(\theta). \tag{2}$$

Naturalmente se espera que α sea lo más pequeño posible, pues constituye una cota superior para el error de tipo I. Sin embargo, valores demasiado cercanos a

cero podrıían ser inconvenientes debido a que aumentarían la probabilidad de error tipo II a niveles no permisibles.

p-valor: En determinados problemas de pruebas de hipótesis, la hipótesis nula resulta ser rechazada, luego, es frecuente preguntarse si se hizo mediante un rechazo contundente (fuerte evidencia en contra) o si no lo fue, dicha situación hace referencia a la necesidad de un instrumento que permita medir la intensidad de la evidencia en contra de H_0 presente en la información muestral. El concepto de p-valor surge en respuesta a esta cuestión.

El p-valor es la probabilidad, asumiendo la hipótesis nula como cierta, de haber observado un valor del estadıístico de prueba al menos tan extremo como el que se observó. Naturalmente el p-valor depende de la distribución del estadıístico T, bajo el supuesto de que H_0 es cierta, como se muestra en la siguiente expresión:

$$p = P(T > T_{obs} | \theta \in \Theta_0),$$

con T_{obs} el valor especıífico de T observado en el experimento. Siendo así, los pasos resumidos para realizar una prueba de hipótesis general, están dados por:

I Plantear la hipótesis nula y la hipótesis alternativa.

II Seleccionar un nivel de significancia α. El umbral probabilıístico bajo el cual la hipótesis será re- chazada.

III Realizar el proceso de muestreo.

IV Elegir la estadıística de prueba adecuada T.

V Encontrar la distribución de T bajo la hipótesis nula.

VI Calcular la región crıítica o región de rechazo C. La región del espacio muestral en la cual la hipótesis será rechazada. Alternativamente, encontrar el valor observado del estadıístico de prueba T obs de la muestra.

VII Encontrar el valor observado del estadıístico de prueba T obs de la muestra. Alternativamente, calcular el p-valor asociado como la probabilidad, bajo la hipótesis nula, de observar un estadıístico de prueba al menos tan extremo como T_{obs}.

VIII Decidir si rechazar o no H_0 con base en la región C especificada en el paso (VI). Alternativamente, rechazar H_0 si el p-valor obtenido es lo suficientemente pequeño de acuerdo con el nivel de significancia previamente especificado.

Teniendo así el panorama de una hipótesis simple, cabe preguntarse por un caso más realista, donde generalmente nos hacemos más de una pregunta como objeto de estudio de alguna investigación y como resultado, tenemos un conjunto de $m > 1$ hipótesis a evaluar. Es aquí donde introducimos el procedimiento de comparación múltiple.

2.2 Procedimiento de Comparaciones Múltiples

Mencionamos antes, que generalmente la mayoría de estudios tienen por objeto el planteamiento de más de una hipótesis, así, es posible juzgar acerca de un determinado número $m > 1$ de hipótesis nulas $H_{01}, \ldots, H_{0m}$. Luego, es pertinente la realización de un procedimiento que nos permita evaluar la veracidad de una hipótesis general, dadas las hipótesis nulas. Generalmente a estos procedimientos se les conoce como *Procedimiento de Comparaciones Múltiples* o *Prueba de Hipótesis Múltiple*.

Usualmente, cuando se realiza un procedimiento de comparación múltiple, nos preguntamos acerca de la veracidad de nuestra hipótesis general; sin embargo también es natural preguntarnos por las hipótesis que hacen consecuente una afirmación acerca de la hipótesis general. Uno de los métodos generales para plantear un problema de comparación múltiple, lo planteó Dutoit et al (2003), en el cual seguimos un algoritmo con los siguientes pasos:

I Elegir y calcular un estadístico de prueba T_j para cada hipótesis individual j y $j = 1, \ldots, m$

II Aplicar un procedimiento de prueba de hipótesis múltiple para determinar cuáles hipótesis se han de rechazar de manera que se controle de alguna forma específica el error tipo I.

2.2.1 Sobre la extensión del caso simple

Cabe recalcar el hecho de que hacer realizar una prueba de manea simultánea al conjunto de hipótesis $\{H_{01}, \ldots, H_{0m}\}$ no es equivalente a realizar m pruebas individuales entre dos hipótesis H_{0i} y H_{0j} ya que, primero, se necesitarían $\binom{m}{2} = \frac{m(m-1)}{2}$ comparaciones individuales. La segunda razón, y tal vez con mayor relevancia, es la independencia. La razón yace en que no existe un supuesto de independencia entre las hipótesis de la colección, de tal manera que, es posible que puedan existir al menos un par de índices i y j tales que el rechazo de H_{0i} podría influir (positiva o negativamente) en las posibilidades del rechazo de H_{0j}. La falta de independencia es, de hecho, un escenario frecuente en la práctica, por ejemplo, en problemas relacionados con genética (Ver por ejemplo ——) y finanzas, donde existen conjuntos masivos de datos altamente correlacionados. Muchos de los procedimientos clásicos dentro de la metodología de comparaciones múltiples requieren el supuesto de independencia entre las hipótesis. Sin embargo, se han desarrollado métodos que realizan modificaciones a los procedimientos, que resultan ser robustos en su implementación.

Otro aspecto que cabe recalcar en el estudio de comparaciones múltiples, está ligado al efecto de la *multiplicidad*. Es necesario un procedimiento agregado, conocido formalmente como *compensación por multiplicidad*, que busca evitar conclusiones sesgadas basadas en situaciones que ocurren por efectos del azar, como se ilustra en el siguiente ejemplo:

Ejemplo 2.1. (Lanzamiento de monedas) Supóngase que un experimentador desea probar estadísticamente si una moneda determinada está balanceada.

Para ello realiza 10 lanzamientos, de los cuales 9 resultan en cara. Si asumimos como cierta la hipótesis de que la moneda es justa entonces la probabilidad de que se observe un resultado al menos tan extremo como ese, sería de $(10 + 1)(1/2)^{10} = 0.0107$, con lo que podemos concluir que no es razonable asumir que la moneda es justa con base en la información obtenida. Si el experimentador deseara repetir la prueba anterior, pero esta ocasión deseara probar a 100 monedas diferentes, se enfrentaría a una prueba de hipótesis múltiple. Dado que la probabilidad de que una moneda justa caiga al menos 9 veces cara cuando se lanza 10 veces es de 0,0107, el experimentador esperaría que observar un resultado como éste al lanzar 100 monedas justas fuera un evento igual de raro; sin embargo, lo cierto es que observar al menos una de las 100 monedas comportarse de esa manera es un evento muy probable, incluso en el caso en que todas sean justas. En efecto, la probabilidad de que en 100 experimentos con monedas justas, al menos una muestre 9 o más caras en 10 lanzamientos está dada por $1 - (1 - 0,0107)^{100} = 0.6604$, por lo que, aplicar el el criterio anterior para probar la hipótesis de que las 100 monedas son justas constituiría un error importante.

El anterior ejemplo, nos muestra la delicadeza de la multiplicidad al momento de trabajar procedimientos de comparación múltiples, pues conforme el número de hipótesis incrementa, la noción de error se complica de manera creciente. Por ejemplo, si una prueba simple se hace a un 5% de confianza, afirmamos que existe un 95% de probabilidad de que la hipótesis nula sea rechazada incorrectamente. Sin embargo, si se realizan m = 100 pruebas de hipótesis simultáneamente, donde todas son ciertas, el número esperado de rechazos incorrectos es 5, mientras que, si las pruebas son independientes, la probabilidad de rechazar al menos una hipótesis incorrectamente es de $1 - (0, 05)^{100} = 0, 994$. Así, conforme m, el número de hipótesis en prueba, se hace grande, dicha probabilidad se acerca a uno sin importar el nivel de significancia en consideración. En este contexto, el error de rechazar una hipótesis nula que es cierta se conoce comúnmente como *falso positivo* o error de tipo I como en el caso de las pruebas de hipótesis simples. Existen en la literatura distintas técnicas para controlar el número de falsos positivos asociados con una prueba de hipótesis múltiple; se pretende ofrecer un panorama general de las técnicas más relevantes en las secciones siguientes, un resumen detallado puede consultarse en Dudoit et al. (2003) y Farcomeni (2008).

2.2.2 Sobre el error

Una vez introducimos el caso múltiple, reemplazamos la única hipótesis de trabajo H_0 , por una colección de hipótesis H_{0j} para $j = 1, 2, \ldots, m$,luego , el concepto de error se vuelve naturalmente más complejo. Bajo este panorama, el interés se generaliza de la probabilidad de rechazar incorrectamente cada hipótesis partıícular al número de hipótesis rechazadas incorrectamente que denotaremos por R. Para introducir los errores en los procedimientos de comparación múltiple, usamos la notación planteada por Benjamini-Hochberg(1995), y la resumimos en la siguiente tabla:

	Hipótesis No Rechazadas	Hipótesis Rechazadas	Total
Hipótesis Verdaderas	U	V	m_0
Hipótesis Falsas	K	S	m_1
	$m - R$	R	M

Donde:

- m es el total de hipótesis realizadas.
- m_0 es el número de hipótesis nulas verdaderas, parámetro desconocido.
- $m - m_0$ es el número de verdaderas hipótesis alternativas.
- V es el número de falsos positivos (error tipo I) (también conocido como *falso descubrimiento*).
- S es el número de verdaderos positivos (conocido como *descubrimiento verdadero*).
- K es el número de falsos negativos (error tipo II).
- U es el número de verdaderos negativos.
- $R = V + S$ es el número de hipótesis nulas rechazadas (conocido como *descubrimientos*, independientemente de si son verdaderos o falsos)

Naturalmente, un investigador estará interesado en minimizar V y K pero, al igual como sucede en el caso univariado, realizar esto simultáneamente es imposible. Por tanto, todo procedimiento estándar de prueba de hipótesis múltiple tendrá como prioridad controlar V o una función de V a algún nivel específico de confianza α. La cantidad en función de V que es de interés controlar recibe el nombre de tasa de error y existe en la literatura en varias formas que ofrecen distintos grados de control a distintos grados de complejidad. Como se definió anteriormente en el caso univariado, el control del error tipo I viene dado por α. Sin embargo, la extensión a las pruebas múltiples viene acompañada de distintas tasas de errores, las cuales presentamos.

Tasa de Error por Comparación (PCER): Consiste de el valor esperado de errores de tipo I dividido entre el número total de hipótesis:

$$\text{PCER} = \frac{\text{E}(V)}{m}$$

la tasa de error por comparación, fue creada con el fin de hacer la analogía del nivel de significancia α de las pruebas individuales, en comparaciones múltiples. Para ver esto, supongamos, por ejemplo, que todas las hipótesis son ciertas y que se prueban individualmente a un nivel de significancia común α. Luego, V es una variable aleatoria cuya distribución es binomial con probabilidad de éxito dada por la probabilidad de rechazar una hipótesis cierta, que es precisamente α. Por tanto, $\text{PCER} = \text{E}(V)/m = m\alpha/m = \alpha$. En general, si m hipótesis

son probadas a un nivel α de significancia, entonces el PCER será siempre α, implicando que no dependa del número de hipótesis realizadas. Esto presenta un problema, ya que se ignora la multiplicidad del problema.

Tasa de Error Global (FWER): Es la probabilidad de cometer uno o más errores de tipo I:

$$\text{FWER} = P(V \geq 1)$$

o equivalentemente,

$$\text{FWER} = P(V > 0) = 1 - P(V = 0)$$

Hochberg and Tamhane (1987) define el término familia como toda colección de inferencias estadísticas para las cuales hace sentido tomar una forma de error combinado o global. La FWER recibe su nombre de una idea similar en la cual es necesario resumir el error global de las pruebas que intervienen en una MCP mediante una cantidad así denominada.

En el Ejemplo 2.1, se presentó de manera natural sin nombre, mostrándonos la necesidad de aplicar procedimientos de control para el mismo.

Tasa de Falsos Descubrimientos (FDR). No satisfechos con los procedimientos para controlar el FWER y PCER, Benjamini and Hochberg (1995) introdujeron una tasa de error que consiste de la proporción esperada de errores entre las hipótesis rechazadas. Formalmente, si definimos la variable aleatoria Q como:

$$Q = \begin{cases} \frac{V}{R}, & R > 0 \\ 0, & R = 0 \end{cases}$$

3 Procedimientos de Control

Si suponemos que FWER$\leq \alpha$, decimos que la probabilidad de cometer un error tipo I está controlada por un nivel α. Un proceso controla el FWER *débilmente* si el control del FWER a un nivel α, es garantizado sólo cuando todas las hipótesis nulas son ciertas. Esto es, cuando $m_0 = m$, esto implica que la hipótesis general H_0 es cierta. Por otro lado, decimos que un procedimiento controla *fuertemente*, si el control del FWER a un nivel α independientemente de la configuración de hipótesis falsas o verdaderas.

Algunos de los procedimientos recientes contolan fuertemente el FWER. Presentamos algunos.

Procedimiento de Bonferroni: Sea $\{H_{01}, \ldots, H_{0m}\}$ una familia de hipótesis; sea p_i el p-valor correspondiente a la hipótesis H_{0i}. Procedemos a rechazar la hipótesis H_{0i}, si $p_i \leq \frac{\alpha}{m}$. El control puede ser probado a través de la *desigualdad de Boole:*

$$\text{FWER} = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha.$$

Procedimiento de Šidák: Dadas m hipótesis nulas y un nivel α , cada hipótesis es rechazada si el p-valor es menor que $\alpha_{SID} = 1 - (1-\alpha)^{\frac{1}{m}}$. Este

procedimiento produce un FWER exactamente de α cuando las pruebas son independientes dos a dos y las hipótesis nulas son verdaderas.Es un poco menos conservadora que la de Bonferroni, pero sólo un poco. Por ejemplo, para $\alpha = 0.05$ y $m = 10$, el nivel de ajuste por Bonferroni es 0.005 mientras que el ajuste de Šidák es 0.005116 aproximadamente.