Tema 12. Introducción al contraste de hipótesis

Primera parte tomada del tema 8 de: PÉREZ JUSTE, R., GALÁN GONZÁLEZ, A. y QUINTANAL DÍAZ, J. (2011) *Métodos y Diseños de Investigación en Educación.* Madrid: UNED©, redactado y ampliado por **Arturo Galán©.**

ÍNDICE

PRESENTACIÓN

En este tema se pretende introducir al lector en los conceptos clave y la lógica de pensamiento de cara a la puesta en práctica e interpretación de un contraste de hipótesis. Aunque la mayoría de los ejemplos se refieren a un diseño de dos grupos independientes, la lógica es aplicable a cualquier diseño de investigación. La realización del contraste de hipótesis presupone el cumplimiento de determinados supuestos rigurosos que en muchas ocasiones no pueden cumplirse en la investigación educativa. Tras el estudio del tema, el lector debe ser capaz de entender los conceptos básicos para enfrentarse con la dinámica del contraste de hipótesis y debe saber ser crítico a la hora de interpretar tanto los resultados de sus propios estudios como los que puedan ser revisados en la literatura científica para la formación permanente del profesional de la educación.

1. INTRODUCCIÓN

Hemos visto en el tema anterior el proceso para establecer un intervalo de confianza para estimar el parámetro desconocido de una población a partir de la información conocida de una muestra. Así, hemos analizado, por ejemplo, cómo es posible estimar el intervalo confidencial de la media poblacional en CI con los datos obtenidos en una muestra y hemos visto que también es posible compararlo con una media poblacional de CI dada.

Sin embargo, la media muestral puede ser utilizada de otra forma: para hacer inferencias sobre los valores probables de una media poblacional desconocida, μ. Los valores de la media muestral, M, pueden ser usados para probar la hipótesis sobre un valor específico de una media poblacional desconocida a través del uso de la prueba de significación de la hipótesis nula (PSHN). Veremos posteriormente la lógica de este proceso.

El alumno de Pedagogía debe tener, en este momento, una idea bastante certera de lo que es una hipótesis y de su relación directa con el problema de investigación. De acuerdo con Salkind (2010), a un buen científico le gustaría poder decir que si el método A es mejor que el método B, esto es cierto para siempre y para cualquier persona en el universo. Sin embargo, es improbable ser capaz de decir algo así, ya que

exigiría un enorme gasto de dinero (¡probarlo en toda la población!) y de tiempo. Por ello, lo que hacen los investigadores es seleccionar una muestra de la población y probar nuestra hipótesis sobre los métodos A y B.

El proceso de selección de la muestra y sus condicionantes (representatividad y tamaño) es decisivo en las decisiones e interpretación de los resultados tras el contraste de hipótesis.

2. Lógica de la prueba de significación de la hipótesis nula.

Sabemos que la hipótesis nula representa la *no relación* entre las variables que estamos estudiando (variable independiente o VI y variable dependiente o VD). Dicho de otra forma, es la hipótesis de *no diferencias* (no hay diferencias en los niveles de la VD (rendimiento, actitud, inteligencia, presión arterial, razonamiento espacial…) en función de la VI (sexo, lugar de nacimiento, fumador, raza, nivel de estudios de los padres…). En otras palabras, en un diseño experimental, significa que la VI no produce ningún efecto en la VD.

La hipótesis nula tiene dos propósitos básicos (Salkind, 2010: 164):

a) Sirve como punto de partida cuando no tenemos conocimiento o no hay razones para creer que existen diferencias entre los grupos que estamos comparando. Por ejemplo, estamos estudiando la actitud de los alumnos de Pedagogía hacia la Estadística y queremos contrastar si hay diferencias entre hombres y mujeres. A falta de otras evidencias, nuestra hipótesis será que no existen diferencias entre las actitudes hacia la Estadística entre los alumnos y las alumnas. Hasta que podamos probar que hay diferencias, debemos asumir, como punto de partida, que no las hay. Y aún más: asumimos que, si encontramos alguna diferencia, esta diferencia es *casual*, se obtuvo por simple casualidad, por azar. El azar explica lo que no tiene otra explicación razonable (por ejemplo, por qué yo obtuve un póker de ases y mi rival sólo una pareja de sietes).
b) Es un punto de referencia para comparar los resultados obtenidos y deducir si las diferencias observadas pueden ser atribuidas a algún factor distinto a la casualidad. De este modo, como se explicó en el tema anterior, la hipótesis nula

ayuda a definir un intervalo en el que cualquier diferencia observada entre grupos puede atribuirse a la casualidad o azar y otro intervalo de valores en el que dicha diferencia quizás se deba a otro factor distinto del azar, como la manipulación de otra variable (VI) que esté ejerciendo un efecto sobre la variable medida (VD) en los grupos que estamos comparando.

Decíamos en la introducción que los valores de la media muestral, M, pueden ser usados para probar la hipótesis sobre un valor específico de una media poblacional desconocida a través del uso de la prueba de significación de la hipótesis nula (PSHN). ¿Cuál es la lógica de este proceso? Warner (2008) indica que hay que seguir los siguientes pasos:

1. El primer paso de este proceso es que el investigador hace una conjetura sobre un valor específico de un parámetro (digamos μ) para una población de interés. Por ejemplo, podríamos conjeturar que la media de velocidad en autovía es de 120 km/h, el límite legal. H_0: $\mu = \mu_{hip}$ o, de otra forma, $\mu - \mu_{hip} = 0$, donde hipotetizamos que H_0: $\mu = 120$. En este caso, el "efecto" que el investigador trata de detectar es la diferencia entre la media poblacional desconocida, μ, y la media poblacional hipotetizada.
2. El investigador selecciona una muestra aleatoria de la población (en este caso, tomando controles de velocidad en distintos tramos de autovías y calculando la media aritmética, digamos M= 132 Km/h).
3. Ahora, el investigador comparará la media observada (M= 132 Km/h) con la media hipotetizada de la población (μ= 120) y se hará la siguiente pregunta: Si la media de la población es realmente de 120 Km/h, ¿puede considerarse la media obtenida en la muestra (M = 132 Km/h) como un resultado probable o improbable? El intervalo de valores establecidos como *improbables* viene dado por el nivel elegido del nivel de significación (alfa, α) y por la elección del tipo de contraste (de una cola o de dos colas, como veremos más tarde). En este momento y de modo general, podemos decir que los valores cuya probabilidad de aparición es igual o menor de 0.05 (es decir, que aparecen 5 veces o menos de cada 100 sucesos) cuando la H_0 es verdadera, son considerados *improbables*.
4. Evaluación de la probabilidad asociada al valor empírico obtenido, supuesta H_0 verdadera. Hemos adelantado en el capítulo anterior cómo afrontar este juicio.

Nos basamos en la distribución muestral teórica del estadístico en cuestión. En el caso de la media, para muestras grandes, sabemos que su distribución muestral puede ser la distribución normal. Por tanto, para evaluar cuán lejos está M del valor μ_{hip} nos basamos en el error típico, es decir, medimos dicha distancia en unidades de error típico. En este caso, la distancia entre M y μ_{hip} es llamada *z* o puntuación típica estandarizada, cuya distribución muestral es la distribución normal, de modo que podemos conocer las áreas o probabilidad de obtener cualquier valor de *z* (usando las tablas o un programa informático), mediante la fórmula general:

$$z = \frac{M - \mu_{hip}}{\sigma_M}$$

En el ejemplo que estamos poniendo, *M* es la media aritmética de una muestra aleatoria, pero en los diseños de dos grupos, por ejemplo, *M* será la diferencia entre las medias de los dos grupos sometidos a contraste, mientras que μ_{hip} será la media hipotetizada según la distribución muestral de diferencia de medias, que no es otra sino H_0: $\mu_A - \mu_B = 0$. El error típico, a su vez, tendrá que ser sustituido por el error típico de diferencia de medias. En línea con el ejemplo anterior, podríamos estar comparando aquí la media en velocidad de la autovía de A Coruña frente a la obtenida en la autovía de Extremadura. O, si ponemos un ejemplo educativo, se trataría de comparar la media en rendimiento lector obtenida con un método experimental con la media obtenida con el método tradicional.

Una vez obtenido el valor de *z*, dado que conocemos su distribución muestral, es fácil responder a la pregunta que planteábamos en el punto 3: Este valor empírico de *z*, que estandariza la distancia entre nuestro valor muestral y el hipotetizado según H_0, ¿puede considerarse probable o improbable? O, dicho de otra forma, el valor muestral obtenido, una vez estandarizado (convertido en una puntuación *z*) y situado en la distribución muestral que incluye infinitos posibles valores de *M*, ¿está cerca o lejos del valor hipotetizado conforme H_0? y, por tanto, ¿cuál es la probabilidad de obtener un valor como el obtenido si es cierta la hipótesis nula? Cuanto más lejos esté el valor de *M* del valor hipotetizado, más grande será el valor de *z* y, en consecuencia, menor será su

probabilidad de aparición si H_0 es verdadera. Como sabemos, el valor de *z* nos indica dicha distancia en número de errores típicos.

En definitiva, la idea básica de la PSHN es que el investigador asume un valor para la media poblacional desconocida, μ. Entonces obtiene una media muestral *M* y la evalúa conforme a la distribución de valores de *M* que cabría esperar si la H_0 es verdadera. El investigador debe tomar la decisión si **rechazar o no rechazar H_0** como inadmisible**,** dado el valor obtenido de *M*. Cuando el valor de *M* es un valor que puede ocurrir probablemente por casualidad cuando H_0 es verdadera, entonces la decisión es no rechazar H_0. Por el contrario, si *M* es improbable que ocurra por casualidad o azar cuando H_0 es verdadera, entonces el investigador puede decidir rechazar H_0. ¿Cuándo estimamos que es improbable que ocurra por azar? Dependerá del valor de α que hayamos establecido *a priori* (y nunca después de haber calculado *z* y conocer su *p* asociada). En este sentido, siguiendo a Siegel (1988), la condición de rechazo es la siguiente:

RECHAZO H_0 si p (t, z, F...) ≤ α

Sin embargo, esta forma de actuar ha tenido en los últimos años serias críticas (nos detendremos en ellas posteriormente), especialmente porque en la investigación real no se cumplen los supuestos que rigen la lógica de la PSHN, en especial, la ausencia de una población claramente definida, la selección de una muestra aleatoria de la misma (en la mayoría de los casos se trabaja con muestras incidentales o de conveniencia) y la realización de una única prueba de significación. En este sentido, en el campo de la Educación, conviene plantearse los resultados obtenidos más como "evidencia acumulativa" que como una prueba irrefutable para tomar decisiones. En esta línea, como veremos después, será de gran ayuda recurrir a índices complementarios como la magnitud del efecto.

3. Nivel de significación estadística y errores Tipo I y Tipo II

El nivel de significación, α, es lo que se conoce como error tipo I: la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera. Cuando tomamos una decisión sobre rechazar o no la hipótesis nula, podemos obtener cuatro resultados posibles, como vemos en la tabla siguiente:

		Lo que de hecho sucede en la población	
		H_0 es verdadera	H_0 es falsa
Decisión del investigador	Rechazar H_0	Error tipo I (α): nivel de significación	Decisión correcta (1- β): potencia estadística
	No rechazar H_0	Decisión correcta (1- α)	Error tipo II (β)

Tabla 1. Resultados en la decisión estadística

El error tipo I nos llevaría a afirmar, por ejemplo, que un medicamento contra el cáncer es eficaz cuando en realidad no lo es o, en nuestro ámbito, que un método de enseñanza de las matemáticas es mejor que otro cuando lo cierto es que no hay diferencias en los resultados de aprendizaje que producen. Dicho de otra forma, estaríamos afirmando que la diferencia de medias entre el grupo experimental y el grupo de control es lo suficientemente grande como para considerar que los grupos provienen de poblaciones con parámetro media (μ) diferente, cuando la realidad es que la diferencia obtenida entre las medias es aleatoria, debida a errores de muestreo, y ambas muestras pertenecen a la misma población.

El error tipo II (β), por su parte, nos llevaría a afirmar que el medicamento no es eficaz, cuando en realidad sí lo es. O nos llevaría a concluir que el nuevo método de enseñanza de las matemáticas no eleva significativamente el rendimiento en matemáticas de los alumnos, cuando en realidad produce mejores resultados que el método tradicional. Por tanto, el error tipo II coincide con la probabilidad de no rechazar la Ho cuando ésta es realmente falsa.

El riesgo de cometer error tipo II depende de varios factores. El tamaño de la muestra es uno de los más importantes, junto con el verdadero tamaño de efecto en la población y el propio valor de α. Cuanto más grandes sean cada uno de estos tres valores, manteniendo los otros constantes, más pequeño será el error tipo II.

En contraposición, la *potencia estadística* (1-β) se define como la probabilidad de rechazar la hipótesis nula cuando ésta es realmente falsa (decisión correcta). Es deseable alcanzar una potencia de la prueba de 0.80 (es decir, $\beta \leq 0.20$). La potencia de la prueba se utiliza, entre otras cosas, para conocer el tamaño que tendría que tener nuestra muestra si queremos alcanzar un determinado valor de β, dado α y un valor previsto para el tamaño del efecto[1]. Los factores mencionados que hacen reducir el error tipo II, son a su vez los que permiten elevar la potencia estadística.

Los investigadores desean mantener ambos tipos de errores tan bajos como sea posible. El valor teórico del error tipo I lo establece el investigador, siendo los valores convencionales utilizados de forma más frecuente 0.01 y 0.05. Cuando obtenemos un valor de nuestro estadístico de contraste (t, z, F…) que estandariza las diferencias empíricas encontradas (por ejemplo, la diferencia de medidas entre dos grupos que estamos comparando), podemos saber –gracias a la distribución muestral de dicho estadístico conforme H_0– la probabilidad de obtener por efecto del azar una diferencia tan grande como la obtenida. Una probabilidad tan pequeña o inferior a α nos estaría indicando que sería muy improbable que dicha diferencia sea aleatoria (pero podría darse el caso, y por eso se trata de un error posible, aunque improbable). Todo esto es cierto en tanto se cumplan los rigurosos supuestos para aplicar la PSHN. Si se violan dichos supuestos (lo que sucede con frecuencia en la investigación educativa), entonces el valor del error tipo I puede ser mayor que el definido por el investigador.

La decisión sobre el valor que debemos dar a α no siempre es fácil. Dependerá en buena parte del tipo de variables que estén en juego. No es lo mismo hablar de un medicamento que puede salvar miles de vidas y que implica el gasto de millones de dólares en su producción, que someter a contraste el efecto de una técnica de estudio en la motivación de los estudiantes. En todo caso, es muy poco frecuente utilizar valores superiores a 0.05, siendo una práctica común utilizar el valor exacto de *p*, como se explicará posteriormente.

4. Los pasos del contraste de hipótesis

[1] Para hacernos una idea del funcionamiento de la potencia, pueden consultarse tablas o páginas como la del prof. Soper http://www.danielsoper.com/statcalc/calc01.aspx

Hemos visto en el tema 2 el proceso y la lógica de la investigación científica. Se parte de la definición de un problema científico. Tras la necesaria revisión del estado de la cuestión (consulta de la bibliografía científica más relevante al respecto), el investigador formula una hipótesis como solución tentativa al problema. Por tanto, el paso decisivo consiste ahora en validar la hipótesis. En el paradigma cualitativo utilizamos la Estadística para tal fin, realizando lo que se denomina *contraste estadístico de hipótesis*. Acabamos de introducir el concepto de *prueba de significación de la hipótesis nula* (PSHN) y hemos comentado la necesidad de cumplir con determinados supuestos. Pues bien, veremos ahora cuál es el proceso básico que deberemos seguir con rigor si queremos llegar a unos resultados fiables e interpretables que nos permitan llegar a validar nuestra hipótesis.

4.1. Selección de una muestra aleatoria

Retomemos el ejemplo del capítulo 11 de Pérez Juste *et al.* (2009). Si estamos estudiando el efecto de un programa para mejorar el rendimiento lecto-escritor de los sujetos, podemos realizar un diseño de dos grupos: a un grupo (grupo experimental) le aplicamos el programa y después le pasamos una prueba de rendimiento lecto-escritor. Al otro grupo (grupo de control) no se le aplica el programa pero se le pasa también la prueba de rendimiento. Supongamos que la población de referencia son los alumnos españoles que cursan tercer curso de educación primaria (no olvidemos que para seleccionar una muestra aleatoria, debemos previamente haber definido con claridad la población). Pues bien, el muestreo aleatorio exige que seleccionemos aleatoriamente de esa población (con un bombo como el de la lotería o, como suele hacerse, con un programa informático) dos muestras de tamaño n y, posteriormente, asignar aleatoriamente (valdría utilizar una moneda) los tratamientos a los grupos. Una vez aplicados los programas a comparar (tratamientos), obtendríamos las puntuaciones de cada sujeto en el test de rendimiento verbal, a partir de las cuales se calcularán las medias aritméticas de cada grupo.

4.2. Formulación de las hipótesis estadísticas

Le corresponde ahora al investigador trasladar su hipótesis sustantiva o textual a hipótesis estadísticas. Supongamos que la hipótesis sustantiva se redacto de la siguiente forma: Con el programa de enseñanza *García* (tratamiento experimental) conseguiremos mejores resultados en rendimiento lecto-escritor que con el programa tradicional.

a) Formulación de la hipótesis nula (H_0)

En primer lugar formulamos la hipótesis que se somete a contraste, esto es, la hipótesis nula. La H_0, como dicen Johnson and Christensen (2008: 503), afirma: "no hay efecto presente" (efecto de la variable independiente sobre la variable dependiente). La hipótesis nula plantea la NO existencia de diferencias entre los grupos sometidos a comparación. Podríamos formularla así: *No existen diferencias estadísticamente significativas entre las medias aritméticas de los grupos en rendimiento lecto-escritor en función del programa utilizado.* Simbólicamente se expresa así:

H_0: $\mu_{Exp.} - \mu_{Cont.} = 0$ o también H_0: $\mu_{Exp.} = \mu_{Cont.}$

Es decir, las diferencias entre las medias aritméticas de los grupos experimental y control son estadísticamente igual a cero, por lo que las diferencias empíricas que existan entre las medias de las muestras se deben al azar; los valores paramétricos son iguales, las dos muestras pertenecen a la misma población.

b) Formulación de la hipótesis alternativa (H_1)

La hipótesis alternativa es llamada también la hipótesis del investigador. Dependiendo de la formulación de la hipótesis sustantiva, la H_1 podrá ser unilateral (también llamada unidireccional o de una cola) o bilateral (bidireccional o de dos colas). Cuando el investigador tiene fundamentos suficientes como para hipotetizar qué grupo de los que estamos comparando obtendrá una media superior, entonces la hipótesis y su posterior contraste será unilateral. En nuestro ejemplo, el investigador tiene razones suficientes (literatura de investigación, evidencia empírica, experiencia práctica) como para pensar que con el método *García* se obtendrán mejores resultados en lecto-escritura que con el tradicional. Por tanto formulará así la hipótesis:

H_0: $\mu_{Exp.} - \mu_{Cont.} > 0$ o también H_0: $\mu_{Exp.} > \mu_{Cont.}$

Es decir, espera que la media en rendimiento del grupo sometido al tratamiento experimental sea superior a la del grupo de control. Como veremos posteriormente, el contraste unilateral puede ser derecho o izquierdo. Cuando la diferencia esperada sea positiva, el contraste será unilateral derecho (y la región de rechazo de la H_0 se encontrará en el extremo derecho de la distribución muestral, en torno a sus valores altos). Cuando la diferencia esperada sea negativa (por ejemplo, se espera que el tratamiento experimental reduzca los niveles de ansiedad), entonces tendremos una hipótesis unilateral izquierda (y la región de rechazo de la H_0 se encontrará en el extremo izquierdo de la distribución muestral, en torno a sus valores bajos).

Sin embargo, cuando no tenemos apoyo como para pensar que la media de un grupo será superior al del otro, es decir, cuando esperamos que haya diferencias entre los grupos pero no sabemos a favor de cuál, entonces procede formular una hipótesis bilateral, que simbólicamente se expresará así:

H_0: $\mu_{Exp.} - \mu_{Cont.} \neq 0$ o también H_0: $\mu_{Exp} \neq \mu_{Cont.}$

Textualmente, podríamos formularlo así: *Existen diferencias estadísticamente significativas entre las medias aritméticas de los grupos en rendimiento lecto-escritor en función del programa utilizado.*

El contraste bilateral es más exigente que el unilateral, ya que implica repartir la probabilidad de error entre las dos colas de la distribución, haciendo dividir por dos el valor de alfa.

4.3. Elección del valor de alfa o nivel de significación

Ya nos hemos referido a esta cuestión anteriormente. Se habla del valor nominal de alfa porque es el investigador el que lo "nomina", el que establece este valor como umbral para emitir un juicio sobre la significación estadística de las diferencias. Aunque los

valores más utilizados son 0.05 y 0.01, el investigador, por las características de su estudio y sus repercusiones (medicamentos, decisiones que implican grandes recursos y grandes cambios, etc.) puede necesitar reducir al máximo el error tipo I y establecer valores inferiores, como $\alpha = 0.001$. También puede suceder que en una investigación exploratoria con muestras pequeñas, el investigador decida utilizar un $\alpha = 0.1$.

4.4. Determinación de la distribución muestral y la región de rechazo

Este es uno de los puntos clave en la comprensión y acierto en la decisión estadística. Debemos recordar que estamos en un paso (contraste estadístico de hipótesis) dentro del proceso general de investigación, que se generó con la definición de un problema científico. En un paso previo al contraste estadístico, habremos apuntado el diseño de investigación que, en este caso, si hubiéramos cumplido la selección y asignación aleatoria, podría tratarse de un diseño experimental de dos grupos. En este tipo de diseños, para muestras grandes, suponiendo que hemos comprobado los supuestos para aplicar pruebas paramétricas, sabemos que la prueba estadística de contraste es la *z* o la t de student. Supongamos que elegimos la z. ¿Cómo se define la distribución muestral del estadístico z? Ya sabemos que el estadístico z responde a una distribución normal. Sabemos también que en un contraste de dos grupos, el estadístico z nos estandariza las diferencias entre las medias de los grupos con respecto a una diferencia igual a cero, es decir, respecto a la diferencia de medias hipotetizada en la H_0. Si nosotros dibujamos la distribución normal conforme a H_0, esta estaría formada por los infinitos valores de

$$\overline{X}_{exp.} - \overline{X}_{cont.}$$

obtenidos en infinitas muestras aleatorias extraídas de la misma población. Esta distribución toma la forma de la curva normal. Si la H_0 es verdadera, la mayoría de las diferencias serán igual o próximas a cero, disminuyendo la altura de la curva (probabilidad de aparición de cada valor) según nos alejemos del valor central que sería el cero. Así obtendríamos diferencias empíricas por efecto del azar de -0.5, de 0.9, de -1,3, de 2,8, etc., tanto más improbables cuanto más grande sea el valor de la diferencia. Se trata, en definitiva, de una distribución muestral que indica la *no existencia de diferencias estadísticamente significativas*. Es decir, es una distribución muestral que

indica que todas las diferencias encontradas entre dos sucesos se deben simplemente al azar.

Como cualquier distribución de frecuencias, la distribución muestral tiene una desviación típica (ver tema anterior) que se denomina error típico, δ. Podremos saber, entonces, cuántas desviaciones típicas se aleja nuestra puntuación empírica (diferencia de medias) de la media de la distribución (que no es más que una diferencia de medias igual a cero).

De este modo, el estadístico de contraste, z en este caso, nos indica cuántos errores típicos se desvía nuestra diferencia de medias de una diferencia de medias igual a cero. Cuando este valor es tan grande que su probabilidad de ocurrencia es igual o menor que el alfa definido por el investigador, entonces diremos que es lo suficientemente improbable como para considerar que dicha diferencia no es aleatoria, sino que se debe a otro factor distinto del azar y que atribuimos generalmente a la variable independiente, en este caso al programa *García*.

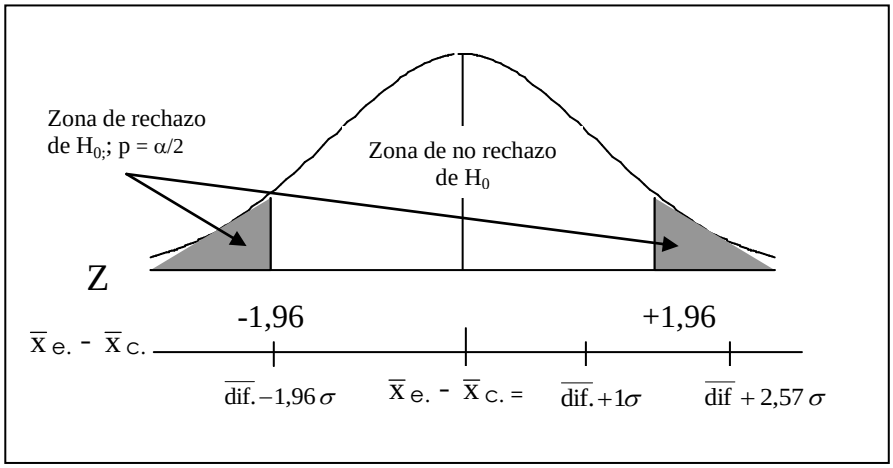


Gráfico 1: Zona de aceptación y rechazo de H_0 en un contraste bilateral ($\mu_E - \mu_C \neq 0$)

En consecuencia, la región de rechazo de H_0 estará formada por todos aquellos valores del estadístico de contraste cuya probabilidad de aparición asociada sea menor o igual que alfa (o α/2 si el contraste es bilateral). De nuevo,

RECHAZO H_0 si p(t, *z*, F...) ≤ α

Esta es la esencia del contraste estadístico de hipótesis. Algunos autores, para tomar la decisión estadística, comparan el valor del estadístico empírico de contraste (el que obtenemos al aplicar la fórmula de z, t, F, χ^2, etc. con el llamado valor crítico del estadístico, que no es más que el valor correspondiente al nivel de significación. Como vemos en el gráfico, para un contraste bilateral con $\alpha = 0.05$ y, por tanto, $\alpha/2 = 0.025$, le corresponde un valor crítico de $z = \pm 1.96$. Esto quiere decir que si nuestro estadístico empírico es mayor que $z = \pm 1.96$, rechazaremos la hipótesis nula.

4.5. Utilización del valor exacto de p

Como habrá entendido el lector, el valor de α se establece a priori, antes de calcular la probabilidad asociada al estadístico de contraste. *p* es precisamente esto, la probabilidad exacta asociada al estadístico de contraste obtenida a posteriori. Nos indica, como ya hemos dicho, la probabilidad de obtener un valor tan extremo o más que el valor observado obtenido al calcular el estadístico de contraste (z, t, F, etc.), *asumiendo que la hipótesis nula es verdadera*. Este valor viene dado por la mayoría de los actuales programas estadísticos utilizados.

Mediante el nivel de significación, llegamos a una decisión binaria: rechazamos o no rechazamos H_0. Con el valor exacto de *p* damos una información mucho más precisa: no sólo si es mayor o menor que el nivel de significación, sino la probabilidad exacta de obtener un valor igual o superior al observado si la H_0 es verdadera. Ello nos permite juzgar si hubiéramos podido rechazar H_0 en caso de que se hubiera elegido un α más exigente. De hecho, muchas revistas científicas del campo de la educación, siguiendo las recomendaciones de la *American Psychological Association* (APA), exigen a los autores informar del valor exacto de *p* en sus artículos. En programas como el SPSS, el valor de *p* aparece bajo la abreviatura **sig.** (significación).

Otra forma frecuente de informar del valor de *p* en artículos científicos en los que se exponen numerosos contrastes en una misma tabla, es utilizar la siguiente nomenclatura:

n.s. ➔ (la diferencia encontrada no es estadísticamente significativa: p > 0,05)
* ➔ p < 0,05 (la diferencia es est. sig. utilizando un α = 0,05)
** ➔ p < 0,01 (la diferencia es est. sig. utilizando un α = 0,01)
*** ➔ p < 0,001 (la diferencia es est. sig. utilizando un α = 0,001)

5. Errores frecuentes en la utilización e interpretación de resultados de la prueba de significación de la hipótesis nula

La PSHN se basa en la asunción de una serie de condiciones:

- Se ha realizado un muestreo aleatorio de N observaciones independientes.
- Las puntuaciones de la variable dependiente son cuantitativas y su nivel de medida es al menos de intervalo.
- Es deseable que la distribución de la variable dependiente sea normal
- El investigador debe seguir rigurosamente los pasos que hemos definido en el epígrafe anterior.
- El investigador debería realizar sólo uno o un número limitado de contrastes en el mismo estudio.

Sin embargo, en la práctica de la investigación educativa, los investigadores violan o se alejan del cumplimiento de dichos supuestos, lo que puede afectar gravemente al error tipo I, elevándolo por encima del alfa nominal.

Por tanto, el investigador educativo debe conocer los supuestos que se deberían cumplir para contrastar la hipótesis nula, así como lo que de hecho se hace en la práctica de la investigación educativa y los efectos de dicha práctica a la hora de interpretar los resultados.

En este sentido, cuando se realizan investigaciones en educación raramente es posible obtener una muestra aleatoria de una población bien definida (pensemos en esta imposibilidad si tenemos la idea de contrastar la eficacia de dos métodos de enseñanza

en un país determinado). Incumplimos, por tanto, una de las condiciones fundamentales para tomar de referencia la distribución muestral conforme a H_0. Esta limitación debe estar bien reflejada en los artículos o informes de investigación, ya que su incumplimiento podría llegar incluso a invalidar el proceso de contraste de hipótesis. Debemos, por tanto, definir tan claramente como sea posible la población de referencia y procurar que la muestra sea representativa de dicha población, si es que queremos realizar inferencias posteriores. Normalmente se habla de una hipotética población de referencia de sujetos similares a los de la muestra de investigación, muestra que generalmente es incidental o de conveniencia. Como ya sabemos, esto limita de forma importante la validez externa.

Otro error frecuente es no tener en cuenta la repercusión de realizar un gran número de contrastes en un mismo estudio. Supongamos, por ejemplo, que hemos aplicado una escala de actitudes de 100 ítems y queremos ver si en cada uno de ellos hay diferencias entre las medias obtenidas entre hombres y mujeres. Suponiendo que en la población no hubiera diferencias entre hombres y mujeres, sólo por el error de muestreo sería esperable, supuesto un $\alpha = 0.05$, que en 5 de esos 100 contrastes obtuviéramos una diferencia de medias significativa cuando en realidad no la hay. Tendríamos entonces lo que se conoce como *riesgo inflado de error tipo I* (Warner, 2008: 96). Existen procedimientos para estimar el nuevo α que deberíamos considerar al aumentar el número de contrastes, como el de Bonferroni.

Para reducir algunas de las complicaciones derivadas de las condiciones reales en las que se puede realizar investigación en el ámbito de la educación, se realizan diversas prácticas. Una de ellas, dirigida a conseguir mayor grado de certeza en las evidencias obtenidas en una línea de investigación es la *replicación del estudio*. Es decir, se hace necesario que otros grupos de investigación, utilizando nuevas muestras y repitiendo el estudio, lleguen a la misma decisión estadística, produciendo así evidencias más conclusivas y reduciendo la posibilidad de error tipo I. También es interesante conocer el método de la *validación cruzada*. Este método supone obtener de nuestra muestra de estudio una *submuestra* utilizando un porcentaje determinado de la muestra inicial.

En esta situación, nuevamente debemos llamar la atención sobre la necesidad de ser conscientes de las graves implicaciones del no cumplimiento de los supuestos para realizar el contraste estadístico de hipótesis, no realizar alegremente dichos contrastes, ser éticos a la hora de hacer nuestros estudios y comunicar los resultados de los mismos

y, finalmente, saber ser críticos con los resultados que leamos en revistas científicas de nuestra disciplina como parte fundamental de nuestra formación profesional a lo largo de la vida.

Por lo que respecta a la interpretación de resultados, aunque habitualmente se habla de aceptar o rechazar la hipótesis nula, en realidad debemos hablar de rechazar o *no rechazar* la H_0. Recordemos que hemos visto anteriormente cuál es la condición para rechazar H_0, no la condición para aceptarla. Por tanto, podemos rechazar la H_0 si cumple dicha condición pero, si no se cumple, simplemente no la rechazamos. Esta distinción de matiz (entre aceptar y no rechazar) es importante porque puede haber muchas razones que expliquen que un resultado ha sido estadísticamente no significativo, cuando en realidad sí lo es (es decir, cuando concluimos que no hay diferencias entre las medias entre dos grupos, puede ser que sí haya tales diferencias en la población). Además del error de muestreo, quizás el ejemplo más claro para entender esto es el tamaño de la muestra. Cuando la muestra utilizada es demasiado pequeña resta potencia estadística a la prueba de contraste y entonces necesitamos diferencias muy grandes para concluir que rechazamos H_0 (cuanto más pequeña es la muestra, más fácil es que la muestra esté sesgada, que no sea representativa de la población, y por tanto habrá que encontrar diferencias muy grandes para poder generalizar que en la población de referencia también hay diferencias). En muchas ocasiones que no hemos rechazado H_0, bastaría haber tenido una muestra más grande para haberla rechazado. Cuanto más grande es el tamaño muestral, más fácil es obtener una diferencia estadísticamente significativa, aunque las diferencias entre las muestras sean pequeñas.

Otra explicación factible es que no se eligieron o no se manipularon bien los niveles de la variable independiente, es decir, volviendo a nuestro ejemplo, significaría que los programas de enseñanza son tan parecidos que no permiten mostrar las diferencias que se habrían obtenido en rendimiento lecto-escritor si el programa experimental no tuviera características en común con el tradicional (puede pensarse también en la cantidad o la dosis de medicamento suministrada a los pacientes para mostrar su efecto real). También podría suceder que estemos midiendo inadecuadamente la variable dependiente (poca fiabilidad y validez) o que sea insensible a los efectos de la variable independiente. En definitiva, concluimos que, en las condiciones en las que hemos realizado nuestra investigación, *no podemos rechazar la hipótesis nula*, lo cual no

quiere decir que sea cierta o aceptemos la hipótesis nula, porque en otras condiciones (mayor tamaño de la muestra, niveles de la VI…) podríamos haber llegado a rechazarla (por eso en Educación es tan importante la replicación de las investigaciones).

Por otra parte, la interpretación de un resultado estadísticamente significativo también debe hacerse con cautela. Deben tenerse en cuenta todos los supuestos ya comentados para aplicar la prueba de significación de la hipótesis nula. Además, deben tenerse en cuenta algunos efectos indeseados que afectan a la validez interna, como la expectativa del investigador de que se cumpla su hipótesis o de que tenga éxito el tratamiento experimental. También debe considerarse la posibilidad de que nuestra variable independiente correlacione altamente con otra variable independiente no tenida en cuenta y que sea esta última, en realidad, la que está produciendo los efectos sobre la variable dependiente que atribuimos a la primera. Como afirma Wagner (2008: 102), "para evaluar si un valor de *p* aporta una información precisa acerca de la magnitud del riesgo de cometer error tipo I, los investigadores necesitan: comprender la lógica de los condicionantes de la PSHN; reconocer los procesos que pueden violar dichos condicionantes en sus estudios; y, finalmente, darse cuenta de cómo dichas violaciones pueden hacer que los valores nominales de *p* sean estimaciones imprecisas del verdadero riesgo de error Tipo I. A este respecto, resulta clarificador el trabajo de López (2003), donde se comentan los análisis alternativos a las pruebas de significación: intervalo de confianza, tamaño del efecto, análisis de la potencia, metanálisis e inferencia bayesiana.

6. El tamaño del efecto

Todo investigador debe ser consciente de que la significatividad estadística nada dice acerca de la magnitud o relevancia práctica o clínica de las diferencias encontradas. Es decir, una diferencia estadísticamente significativa no nos dice que tal diferencia sea *importante*, sino tan sólo que existen diferencias en la población de referencia, aunque dicha diferencia (o correlación, o lo que proceda) sea en la práctica irrelevante.

Por tanto, la significación estadística por sí sola no nos dice gran cosa como para tomar decisiones en la práctica. Además, tampoco permite comparar los resultados entre distintas investigaciones cuando la unidad de medida, la métrica, varía.

Para solucionar este problema, desde hace algunos años se utiliza como indicador complementario a la significación estadística el tamaño del efecto, siendo la *d* de Cohen el índice más utilizado (hay otros como el de Hedges o el de Glass). Este índice muestra el tamaño del efecto como una diferencia tipificada (como una puntuación *z*), donde el numerador es la diferencia de las medias entre los grupos (nos centramos ahora, para su comprensión, en los diseños de dos grupos independientes) y el denominador es una desviación típica, que recibe el nombre de *desviación típica combinada* y que no es más que eso, la combinación de las desviaciones típicas de los grupos que estamos comparando. En definitiva, *d* nos indica cuántas desviaciones típicas se aparta una media aritmética de la otra. La fórmula de la *d* de Cohen es:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{combinada}},$$

donde la desviación típica combinada para diseños de dos grupos independientes es

$$\sigma = \sqrt{\frac{(N_1 \ s_1^2) + (N_2 \ s_2^2)}{N_1 + N_2}}$$

En Ciencias Sociales, los valores sugeridos para interpretar el valor de *d* son los siguientes:

$d \leq .20$ efecto *pequeño*
$d \approx .50$ efecto *moderado*
$d \geq .80$ efecto *grande*

Este valor también se interpreta en los siguientes términos: Si consideramos la media menor como el valor que representa al *sujeto medio del grupo 1* (en el sentido del sujeto prototipo, típico, normal) y la media mayor como el valor que representa al *sujeto medio del grupo 2*, suponiendo que ambos sujetos medios pertenecen a grupos que se distribuyen según la curva normal, si el valor de *d* es 0.5, podríamos decir que el sujeto medio del grupo 2 está media desviación típica por encima que el sujeto medio del grupo 1. Si nos vamos a las tablas de la curva normal, vemos que a una z = 0.5 le corresponde un área hasta la media de *p* = 0.1915, lo que puede interpretarse como que el sujeto medio del grupo dos supera en 19 percentiles (19,15%) al sujeto medio del grupo uno (Morales, 2007: 27).

Otra forma sencilla de valorar el tamaño del efecto es mediante una correlación biserial-puntual. La variable X sería dicotómica, es decir, pertenecer a uno de los dos

grupos que se están comparando (valores posibles 0/1), mientras que la variable Y sería continua, la medida de la variable dependiente. La interpretación es como cualquier coeficiente de correlación de este tipo (cuyo valor varía de 0 a ±1). De esta forma podemos transformar el valor de la *t de student* en otro valor que nos da la idea de la fuerza de la relación entre la variable independiente y la dependiente y que nos permite comparar también los resultados en distintos estudios.

En definitiva, es importante combinar los resultados de la significación estadística y el tamaño del efecto. De esta forma, podremos valorar con más precisión los resultados obtenidos. Es importante ser consciente de que podrían obtenerse resultados aparentemente contradictorios, es decir, rechazar H_0 y obtener un tamaño del efecto poco relevante (algo que puede suceder cuando tenemos muestras muy grandes) o aceptar H_0 y obtener un tamaño del efecto muy relevante (más fácil que suceda cuando las muestras son pequeñas).

7. CONTRASTE DE HIPÓTESIS EN LOS DISEÑOS DE DOS GRUPOS.

Hemos visto hasta el momento la introducción a la inferencia estadística y al contraste de hipótesis. Se ha tratado la comparación entre la estimación de parámetros y el contraste de hipótesis, tanto cuando tenemos un grupo (por ejemplo, para responder a la pregunta ¿será la media de este grupo igual a 20?) como cuando tenemos dos grupos (por ejemplo, al responder a ¿pueden considerarse estadísticamente significativas las diferencias entre las medias de estos dos grupos?). Hemos visto también cómo se formulan las hipótesis estadísticas, la nula y la alternativa.

Pues bien, en la investigación, en muchas ocasiones lo que buscamos es conocer los efectos de una variable, la independiente, sobre otra que recibe su influjo, la dependiente. Recordemos lo que tratábamos en el capítulo 2 sobre la clasificación de variables según un criterio metodológico-experimental:

- *Variable Independiente (V.I.):* Es la característica que el investigador observa o manipula deliberadamente para conocer su relación o influencia sobre la variable dependiente. En una relación causal, la variable independiente es la causa y la dependiente el efecto. La V.I. suele ser el estímulo, el tratamiento o la situación experimental.
- *Variable dependiente (V.D.):* Es la característica cuyas variaciones se espera que se produzcan por el efecto de la V. I. Son observables y medibles. La V.D. suele recibir el nombre de "variable medida".
- *Variables de control, moderadoras o controladas:* Son aquellas que sabemos que tienen influencia sobre la V.D. y controlamos dicha influencia para que los cambios

que se produzcan en la V.D. los podamos atribuir sólo a la V.I. y no a otras variables extrañas.

- *Variables extrañas:* Son aquellas que intervienen en el experimento pero que se escapan al control del investigador (condiciones ambientales, estado físico y anímico...). No podemos desligar su influencia en la V.D. de la que produce la V.I.

En resumidas cuentas, lo que buscamos es saber si es cierta esta relación causa-efecto:

VI → VD

Así, por ejemplo, lo que se desea es saber si un medicamento (VI) es eficaz en la cura del cáncer (VD), una terapia (VI) es eficaz para reducir los niveles de ansiedad (VD) o un método de enseñanza (VI) es eficaz para conseguir un mejor rendimiento en matemáticas (VD). Una forma de resolver este problema es mediante los diseños de dos grupos: A un grupo (llamado grupo experimental) se le aplica el tratamiento o método experimental (VI) y a otro grupo no se le aplica y se utiliza como grupo de comparación (es el grupo de control). Después de aplicar el tratamiento, suponiendo controladas el resto de variables intervinientes, se comparan los grupos en la variable dependiente para ver si la VI produjo o no algún efecto sobre la VD. Para ello, lo que se comparan son las medias de los grupos (estamos simplificando a efectos didácticos). Generalmente, el investigador esperará rechazar la hipótesis nula, ya que confiará en que su tratamiento sea eficaz; sin embargo, deberá verificar su intuición empíricamente.

Acabamos de referirnos a un *diseño* de dos grupos. Un DISEÑO es el plan (esbozo del proyecto), estructura (esquema) y estrategia (recogida y análisis de datos) de investigación concebidos para obtener respuestas a las preguntas de investigación y controlar la varianza (la varianza de la variable dependiente, es decir, qué causas producen las variaciones, las diferencias, en la VD).

Pues bien, muy resumidamente, hay dos tipos básicos de *diseños*:

- **experimentales**: Son aquellos en los que existe manipulación y control de la/s V.I. Es decir, tendremos diseños experimentales cuando se aplica tratamiento, un método, etc., y nosotros, como investigadores, hemos establecido sus características y condiciones (a esto se refiere el término "manipulación de la VI). Además, hay selección aleatoria de los sujetos de la muestra a partir de los sujetos de la población (algo muy complicado en educación; cuando no se cumple esta selección, hablamos de diseños cuasi-experimentales) y selección aleatoria en la asignación de los tratamientos a los grupos (se sortea, por ejemplo, qué grupo será el experimental y cuál el de control).
- **no experimentales, ex-post-facto o correlacionales**: Son aquellos en los que no hay manipulación ni control de la/s V.I. porque ya acontecieron sus manifestaciones o por ser intrínsecamente no manipulables. Por tanto no puede haber grupo experimental y de control, tan sólo comparamos las medias en la VD en función de una variable de clasificación que ya ha acontecido. Por ejemplo, cuando queremos ver las diferencias en rendimiento en función del sexo, en función del nivel familiar o en función del cociente intelectual. Estas variables, o bien son intrínsecamente no manipulables, o tenemos que esperar a que acontezcan para estudiar sus posibles efectos sobre la VD. En estos casos, se hacen inferencias sobre las relaciones entre las V.I. y las V.D. a partir de la variación común de las mismas, por lo que, en sentido estricto, no podemos hablar de relación causa-efecto.

En el capítulo 2 se expusieron los pasos del proceso de investigación, que pueden sintetizarse así:

1. Formulación del problema
2. Revisión bibliográfica del estado de la cuestión
3. Definición de variables
4. Formulación de hipótesis
5. Diseño de la investigación
6. Contraste estadístico de hipótesis:
    - 6.1. Formulación de las hipótesis estadísticas
        - Hipótesis nula: H_0
        - Hipótesis alterna: H_1
    - 6.2. Elección de la prueba estadística adecuada
    - 6.3. Establecimiento del nivel de significación (α)
    - 6.4. Definición de la distribución muestral respecto H_0
    - 6.5. Establecimiento de la región de rechazo de H_0
    - 6.6. Decisión estadística (rechazo o no rechazo de H_0)
7. Decisión práctica (magnitud del efecto y conclusiones generales)

Tabla 1: Fases del proceso de investigación

Como vemos, la investigación siempre se inicia con el planteamiento de un problema al que queremos dar respuesta. En este caso nos referimos a problemas que puedan resolverse mediante una metodología cuantitativa, lo que no quiere decir que otro tipo de problemas que requieren distintas metodologías para su solución no sean científicos. Un ejemplo de problema podría ser el siguiente: ¿Existen diferencias en la motivación hacia el estudio entre los alumnos superdotados y el resto de alumnos de secundaria?

A partir del planteamiento del problema de forma más o menos compleja (pero siempre incluyendo las variables fundamentales del estudio, VI y VD) debemos revisar el estado de la cuestión, es decir, revisar en las bases de datos científicas qué se ha investigado anteriormente sobre nuestro tema, con el fin de aplicar los hallazgos a los que han llegado otros antes que nosotros. A continuación definimos cuáles serán las variables que intervendrán en la investigación y formularemos nuestra hipótesis, es decir, la solución más probable al problema de investigación. Por ejemplo, los alumnos superdotados tienen un nivel de motivación hacia el estudio más bajo que el resto de alumnos. Esta hipótesis, que recibe el nombre de *hipótesis sustantiva*, convenientemente transformada en *hipótesis estadísticas*, será la que contrastemos mediante la prueba de significación de la hipótesis nula (PSHN).

El diseño de investigación, entendido de modo amplio, como dice la famosa definición de Kerlinger (1985), es el plan, la estructura y la estrategia de investigación concebidos para dar respuesta a preguntas de investigación y controlar la varianza. El plan es el esbozo general del proyecto de investigación (incluye desde la formulación de hipótesis hasta el análisis de los datos). La estructura es el esquema de lo que se hará con las variables. La estrategia se refiere a los métodos de recogida y análisis de datos (muestreo, instrumentos, etc.). En sentido más reducido, que es el que indicamos en el esquema del contraste de hipótesis, el diseño se refiere al procedimiento más adecuado para contrastar nuestra hipótesis.

En el ejemplo anterior, estaríamos ante un diseño no experimental de dos grupos con medida postest. En este caso no hay grupo experimental y de control porque se trata de un diseño no experimental, sin posibilidad de manipular la VI que, en este caso, es la presencia o no de *alta capacidad intelectual*.

A partir de aquí se puede proceder al contraste estadístico de hipótesis, cuyos pasos los hemos explicado al comienzo de este capítulo.

Prueba estadística para el contraste de hipótesis

La prueba estadística dependerá del tipo de diseño. En este momento vamos a estudiar el diseño más sencillo, que es el diseño de dos grupos independientes. Es decir, tenemos dos grupos establecidos por la variable de clasificación, que es la VI; decimos que estos dos grupos son independientes porque el pertenecer a un grupo es independiente de pertenecer al otro (a diferencia de lo que sucede en el caso de grupos relacionados[2]).

Una vez que está claro el tipo de diseño, hay que elegir la prueba de contraste más adecuada. Existen dos tipos de pruebas: las paramétricas y las no paramétricas. Las primeras son las más potentes para rechazar H_0, por lo que son las preferibles. Pero para utilizarlas, se deben cumplir los siguientes supuestos (partiendo del supuesto de ***independencia de las observaciones*** que incumplimos casi siempre en Ciencias de la Educación y que se refiere a que se realizó muestreo aleatorio en la selección de las muestras, como ya vimos al hablar de la PSHN):

1. ***Nivel de medida de la Variable Dependiente:*** debe ser de **Intervalo o Razón**. En Ciencias Sociales se acepta también el nivel de *cuasi-intervalo* (por ejemplo, las escalas tipo Likert que indican el grado de acuerdo con respecto a una afirmación: "*Valore de 1 a 6 su grado de acuerdo con las siguientes afirmaciones…*").
2. ***Normalidad de la distribución en la población:*** La distribución de la Variable Dependiente en la población se distribuye según ***la curva normal***. Se comprueba con el test de χ^2 o el de Kolmogorov-Smirnov.
3. ***Homocedasticidad de las varianzas poblacionales:*** Igualdad en las varianzas, es decir, los dos grupos que comparamos tienen varianzas estadísticamente iguales o pertenecen a la misma población de varianza σ^2. Se prueba con **el test F de Snedecor o con la F de Levene.**

Siendo puristas, si no se cumplen estas condiciones no se puede utilizar una prueba paramétrica. En la mayoría de las investigaciones es complicado cumplir todas estas condiciones, sobre todo por los problemas para seleccionar las muestras. Algunas pruebas, como las que vamos a ver aquí, se ven poco afectadas por el incumplimiento de los supuestos 2 y 3, pero hay que saberlo e indicarlo.

[2] Existen también los diseños de dos grupos correlacionados. Se habla de grupos correlacionados cuando al mismo grupo se le ha medido en dos ocasiones diferentes (por ejemplo, en situaciones de pretest-postest) o cuando comparamos las diferencias entre las calificaciones en dos asignaturas distintas de un mismo grupo de sujetos. También se habla de grupos correlacionados cuando las puntuaciones que dan origen a las dos medias aritméticas que se comparan, se han ordenado en función de otra variable que recibe el nombre de variable de bloqueo.

Una vez establecido si debemos utilizar una prueba paramétrica o no paramétrica, debemos saber elegir la prueba adecuada. Esto dependerá de aspectos como el **número de variables independientes, el número de grupos o muestras, el tipo de muestras** (independientes o correlacionadas) y el tamaño de la muestra.

En este caso vamos a ver solamente las pruebas para una VI con dos niveles (es decir, para dos grupos), siendo los grupos independientes y para cualquier tamaño muestral.

Para muestras grandes (N>30) la prueba adecuada es z (Razón crítica).

$$z = \frac{|\overline{X}_1 - \overline{X}_2| - 0}{\sigma_{\overline{X}1-\overline{X}2}}$$

donde

$$\sigma_{\overline{X}1-\overline{X}2} = \sqrt{\left(\frac{\sigma_1^2}{N_1 - 1} + \frac{\sigma_2^2}{N_2 - 1}\right)}$$

El poner en el numerador "-0", se hace a efectos didácticos para entender que lo que estamos comparando es nuestra diferencia empírica de medias aritméticas con la diferencia hipotetizada en la hipótesis nula (H_0: $\mu_1-\mu_2= 0$). Al dividir esta diferencia por el error típico de diferencia de medias, el estadístico de contraste nos indicará cuántos errores típicos se aleja nuestra diferencia de una diferencia igual a 0, y cuál es la probabilidad de obtener tal valor por efecto del azar.

Para muestras <u>grandes o pequeñas</u> podemos utilizar la prueba t de Student.

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right| - 0}{\sqrt{\left(\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

En el caso de dos grupos **correlacionados**, la fórmula general para el contraste de medias es:

$$t = \frac{|\overline{X}_1 - \overline{X}_2| - 0}{\sigma_{\overline{X}1-\overline{X}2}}$$

Donde el error típico de diferencia de medias tiene el cuenta el valor de la correlación (r):

$$\sigma_{\bar{X}1-\bar{X}2} = \sqrt{\left(\frac{\sigma_1^2}{N_1 - 1} + \frac{\sigma_2^2}{N_2 - 1}\right) - 2r\left(\frac{\sigma_1}{\sqrt{N_1 - 1}}\right)\left(\frac{\sigma_2}{\sqrt{N_1 - 1}}\right)}$$

Establecer el nivel de significación

Ya se ha comentado que lo establecerá el investigador. Los valores usuales son: α=0,05; α = 0,01; α = 0,001. También puede expresarse, por ejemplo, como α = .05.

Definición de la distribución muestral

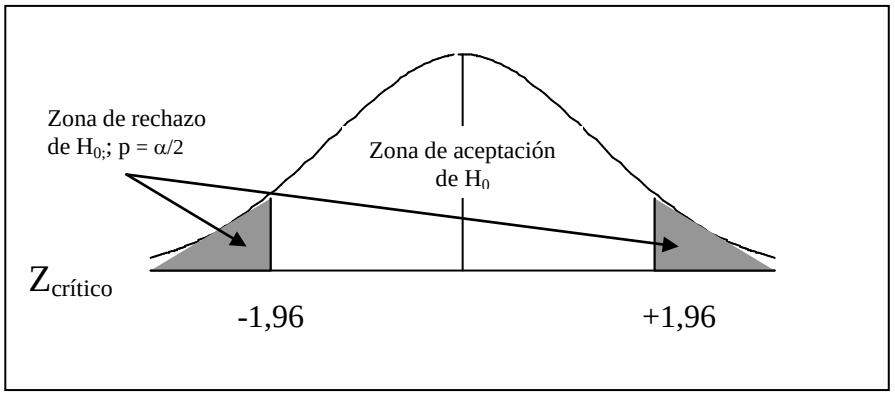
La distribución muestral conforme a la H_0 estaría formada por los infinitos valores del estadístico de contraste obtenidos en infinitas muestras aleatorias de tamaño N extraídas de la misma población.

Definición de la región de rechazo de la distribución muestral

Es aquella parte de la distribución muestral formada por todos los valores del estadístico de contraste cuya probabilidad de aparición asociada sea menor o igual que α. (*o α/2 si el contraste es bilateral).*

Conviene siempre dibujarlo:

Zona de rechazo y no rechazo (aceptación) de H_0 en un contraste bilateral ($\mu_E - \mu_C \neq 0$)



Decisión estadística

Rechazar o No rechazar la Hipótesis Nula **H_0**, la hipótesis de contraste.

Recordemos:

RECHAZO H_0 si p (t, Z, F...) ≤ α

Para ello, calcularemos el valor del estadístico empírico según la fórmula elegida y compararemos la probabilidad asociada a este valor con el valor de α. Por extensión, en las pruebas Z y t, si comparamos el valor del estadístico empírico con el valor crítico correspondiente al nivel de significación, podremos concluir que cuando el valor empírico es mayor que el crítico, se rechaza la hipótesis nula.

Relevancia práctica de las diferencias: Tamaño del efecto

Una vez terminado el contraste estadístico de hipótesis, independientemente de la decisión estadística, debemos calcular el tamaño del efecto para indicar la relevancia práctica de las diferencias. Será un cálculo complementario que nos ayudará a interpretar la *importancia* que debemos atribuir a las diferencias encontradas desde un punto de vista educativo, más allá de lo que se pueda generalizar a la población e independientemente del tamaño de la muestra (aspecto, claro está, que tendremos que considerar al redactar nuestras conclusiones). Para ello podemos utilizar la *d* de Cohen para diseños de dos grupos independientes:

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sigma_{combinada}},$$

donde la σ combinada es:

$$\sigma_{comb.} = \sqrt{\frac{(N_1 \ s_1^2) + (N_2 \ s_2^2)}{N_1 + N_2}}$$

Un ejemplo

Veamos otro ejemplo: estamos estudiando el efecto de un programa para mejorar la capacidad verbal de los sujetos. Para ello, hemos seleccionado aleatoriamente dos grupos de características similares: a un grupo (grupo experimental) le aplicamos el programa y después le pasamos una prueba de rendimiento verbal. Al otro grupo (grupo de control) no se le aplica el programa pero se le pasa también la prueba de rendimiento. De esta forma obtenemos dos medias aritméticas, una por cada grupo, esperando que nuestro programa haya sido eficaz, es decir, que el grupo de sujetos al que hemos aplicado el programa obtenga una media aritmética superior a la del grupo de control.

1. **Diseño de la investigación**

   Se trata de un diseño experimental de dos grupos independientes con medida postest (esto significa que sólo se ha realizado una medición después de aplicar la VI).

   VI: Método
   VD: Rendimiento verbal

2. **Contraste estadístico de hipótesis:**

6.1. Formulación de las hipótesis estadísticas

- Hipótesis nula: $\mathbf{H_0: \mu_E - \mu_C = 0}$ (no existen diferencias estadísticamente significativas entre las medias aritméticas en capacidad verbal de los grupos, en función de la aplicación o no del programa. Dicho de otra forma: las diferencias entre las medias aritméticas de los grupos experimental y control son estadísticamente igual a cero, es decir, las diferencias que existan entre las medias de las muestras se deben al azar; los valores paramétricos son iguales, las dos muestras pertenecen a la misma población).
- Hipótesis alterna: $\mathbf{H_1: \mu_E - \mu_C > 0}$ (Existen diferencias estadísticamente significativas entre las medias aritméticas en capacidad verbal de los grupos, en función de la aplicación o no del programa, favoreciendo al grupo que sigue el método experimental. Dicho de otra forma: la media aritmética del grupo experimental es estadísticamente superior a la media del grupo de control, es decir, existen diferencias estadísticamente significativas entre las medias de los grupos a favor del grupo experimental ➔ tal y como se desprende del enunciado, en el que se especifica la dirección de las diferencias, tenemos que hacer un contraste unilateral).

6.2. Elección de la prueba estadística adecuada

Supongamos que se cumplen los supuestos de independencia, normalidad y homocedasticidad. La VD se ha medido mediante un test estandarizado de capacidad verbal, por lo que su nivel de medida puede considerarse de intervalo. En consencuencia, procede el uso de una prueba paramétrica.

Tenemos una VI, dos grupos, grupos independientes y N pequeña, por lo que seleccionamos la prueba **t de Student**.

Veamos, no obstante, cómo se comprobaría el supuesto de normalidad con SPSS (el de igualdad de varianzas lo veremos en la propia aplicación de la prueba t):

Comprobamos el supuesto de normalidad en cada muestra para $\alpha = .05$:

Muestra baja capacidad:

Prueba de Kolmogorov-Smirnov para una muestra

		Motivacion
N		19
Parámetros normales[a,b]	Media	10,2632
	Desviación típica	4,31846
Diferencias más extremas	Absoluta	,174
	Positiva	,174
	Negativa	-,158
Z de Kolmogorov-Smirnov		,757
Sig. asintót. (bilateral)		,616

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.

Como $p(Z_{k-s}) = 0{,}616 > \alpha$, no se rechaza la H_0 y aceptamos el supuesto de normalidad de la distribución

Muestra alta capacidad:

Prueba de Kolmogorov-Smirnov para una muestra

		Motivacion
N		17
Parámetros normales[a,b]	Media	9,0588
	Desviación típica	3,79919
Diferencias más extremas	Absoluta	,102
	Positiva	,102
	Negativa	-,092
Z de Kolmogorov-Smirnov		,420
Sig. asintót. (bilateral)		,995

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.

Como $p(Z_{k-s}) = 0{,}995 > \alpha$, no se rechaza la H_0 y aceptamos el supuesto de normalidad de la distribución

6.3. Establecimiento del nivel de significación (α)

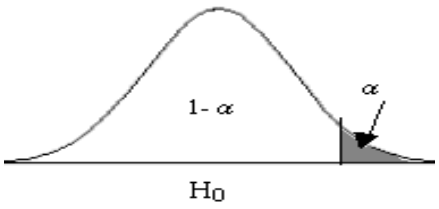
Decidimos un $\alpha = .05$

6.4. Definición de la distribución muestral respecto H_0

La distribución muestral conforme a la H_0 estaría formada por los infinitos valores de t obtenidos en infinitas muestras aleatorias del tamaño dado extraídas de la misma población.

6.5. Establecimiento de la región de rechazo de H_0

Es aquella parte de la distribución muestral formada por todos los valores de t cuya probabilidad de aparición asociada sea menor o igual que $\alpha = .05$.



6.6. Decisión estadística (rechazo o no rechazo de H_0)

RECHAZO H_0 si $p(t) \leq 0{,}05$ (que en este caso, será la $p_{unilateral}(t)$)

Calculamos el resultado con SPSS

Estadísticos de grupo

	Altacapacidad	N	Media	Desviación típ.	Error típ. de la media
Motivacion	,00	19	10,2632	4,31846	,99072
	1,00	17	9,0588	3,79919	,92144

Realizamos ahora el contraste de medias mediante la prueba t. La propia prueba nos indica en primer lugar si se cumple o no el supuesto de homocedasticidad con la F de Levene, con el fin de que utilicemos el valor que corresponda.

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Motivacion	Se han asumido varianzas iguales	1,695	,202	,884	34	,383	1,20433	1,36288	-1,56538	3,97405
	No se han asumido varianzas iguales			,890	33,994	,380	1,20433	1,35299	-1,54529	3,95396

Como p(F) = 0,202 > α, no se rechaza la H_0 y aceptamos el supuesto de homocedasticidad o igualdad de varianzas. Por lo tanto, se cumplen todos los supuestos para utilizar una prueba paramétrica y es correcto utilizar la prueba t.

Analizamos ahora el estadístico de contraste de medias. Vemos que el valor de $t_{empírico}$=0,884, con una probabilidad de ocurrencia asociada de p=0.383. No se cumple en consecuencia la condición de rechazo, ya que $p(t_e)> \alpha$, por lo que podemos considerar que la diferencia de medias obtenida entre los grupos de 1,20 es estadísticamente igual a 0, se ha obtenido por efecto del azar, por lo que no puede afirmarse que haya diferencias en nivel de motivación en función de tener o no tener alta capacidad. No puede confirmarse la hipótesis del investigador con los datos obtenidos en esta muestra, con una probabilidad de cometer error tipo I de 0.05.

7. Decisión práctica (magnitud del efecto y conclusiones generales)

Aplicando la fórmula de Cohen, obtenemos que d= 0,3, es decir, el efecto de la VI sobre la VD es poco relevante[3]. Por tanto, el investigador debería revisar su hipótesis, comprobar la idoneidad de los instrumentos de medida, la selección de los sujetos para asignarles a los grupos de alta capacidad y control (quizás los grupos no eran suficientemente distintos como para que se produjeran diferencias en el caso de que verdaderamente las hubiera). También hay que considerar que el tamaño de las muestras es bastante pequeño. En cualquier caso, con estos datos no debería recomendar utilizar unas técnicas de motivación diferentes con los alumnos de alta capacidad.

[3] Recordemos que partimos en el ejemplo de un diseño no experimental, por lo que no se podría hablar propiamente de relación causa-efecto.

***

Cuando manifestemos las **conclusiones** de una investigación cuantitativa, debemos ser muy cautos. Primero, porque estamos realizando inferencias y, en consecuencia, podemos estar cometiendo un error en nuestras afirmaciones (error cuya probabilidad es conocida). Segundo, porque trabajamos con personas y en situaciones y contextos naturales que hacen difícil el control exhaustivo de variables, por lo que siempre existirán variables extrañas que se escapan a nuestro control. Tercero, porque las posibilidades de generalización de nuestros hallazgos están limitadas a las poblaciones de las que hemos seleccionado las muestras (muchas veces con muestras no aleatorias), por lo que se necesitarán nuevas investigaciones para probar nuestras hipótesis en contextos diferentes. Estas razones, entre otras, hacen necesario en nuestras ciencias la replicación de los experimentos para aumentar la certeza de nuestros hallazgos e ir progresando en el desarrollo de la teoría.

8. RESUMEN

El contraste estadístico de hipótesis es un método, entre otros, para tratar de llegar a evidencias empíricas en el campo de la educación, con el fin de sumar conocimiento para llegar a teorías con un amplio grado de generalidad. El contraste de hipótesis se basa en la prueba de significación de la hipótesis nula, es decir, se parte de que no existen diferencias en la población en aquello que se plantea en la hipótesis nula (por ejemplo, diferencias entre las medias de dos grupos sometidos a distintos métodos o tratamientos). La prueba de significación de la hipótesis nula exige que se cumplan determinados supuestos (población definida, selección aleatoria, contrastes limitados, normalidad…) que muchas veces no pueden cumplirse en la investigación educativa. Por tanto, es necesario conocer dichos supuestos y sus efectos a la hora de interpretar los resultados obtenidos. Por otra parte, el contraste de hipótesis exige seguir unos pasos de forma sistemática y rigurosa (hipótesis, muestra, comparación y decisión), lo que implica conocer los errores que podemos estar cometiendo (error Tipo I y error Tipo II) así como la importancia de la potencia estadística de la prueba utilizada. Estos pasos, entrando ya en su formulación estadística, se concretan en los siguientes: 1. Selección de una muestra aleatoria; 2. Formulación de las hipótesis estadísticas (hipótesis nula y alterna) de acuerdo con la hipótesis sustantiva, reconociendo cuando la alterna será unilateral o bilateral. 3. Seleccionar el nivel de significación. 4. Definir la distribución muestral y la región de rechazo y 5. Decisión estadística: rechazar o no la hipótesis nula. Para un profesional de la educación, obligado a leer artículos científicos para mantenerse actualizado, es esencial saber interpretar también el valor de *p* que suele mostrarse en dichos artículos. Finalmente, la necesidad de ir más allá de la significación estadística para conocer la relevancia práctica de los resultados, nos lleva al cálculo del tamaño del efecto, lo que nos permitirá también la comparación de los resultados de distintos estudios y la acumulación de evidencia científica.

9. BIBLIOGRFÍA

CRESWELL, J. W. (2008). Educational research: planning, conducting, and evaluating quantitative and qualitative research. Upper Saddle River, N.J.: Pearson/Merrill Prentice Hall.

CRESWELL, John W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, Calif.: Sage Publications. 3rd ed.

GALÁN GONZÁLEZ, A. (2003). *La investigación cuantitativa en educación especial.* En J. GONZÁLEZ PÉREZ (Coord.) *Necesidades educativas especiales e intervención psicopedagógica*. Universidad de Alcalá. Madrid.

LÓPEZ GONZÁLEZ, E. (2003). Las pruebas de significación: una polémica abierta. *Bordón*, 55 (2).

MORALES, P. (2008). *Estadística aplicada a las Ciencias Sociales. Contraste de medias.* Universidad Pontificia Comillas, Madrid. [http://www.upcomillas.es/personal/peter/]

SALKIND, Neil J. (2007). Statistics for people who (think They) hate statistics. Sage. California.

TROCHIM, W. M.K. (2009). Research Methods Knowledge Base. [http://www.socialresearchmethods.net/kb/contents.php].

TROCHIM, W. M.K. and Donnelly, J. P. (2009). Research methods knowledge Base, 3e. Cengage Learning, Mason, USA.

WAGNER, Rebecca M. (2008). Applied statistics: from bivariate through multivariate techniques. Sage. California.