UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y SISTEMAS

SOLUCIÓN A UN PROBLEMA DE GRANDES MUESTRAS CON EL EMPLEO DE META-ANÁLISIS

T E S I S

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN ESTADÍSTICA E INVESTIGACIÓN DE OPERACIONES

PRESENTA:
MANUEL GARCÍA MINJARES

DIRECTORA DE TESIS:
DRA. SILVIA RUIZ VELASCO ACOSTA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y SISTEMAS

CIUDAD UNIVERSITARIA, CD. MX., ABRIL 2021

A mi mejor amigo: Jesucristo,

Quién hace posible el querer como el hacer.

Para Adriana, Mel y Luish A.

Mi herencia del Señor.

Agradecimientos

En este momento, sin duda, hay muchas personas a quiénes agradecerles poder llegar a este punto, pero en especial quiero centrarme en tres destacados universitarios con el que he tenido el privilegio de coincidir recientemente, y que han cambiado mi visión y perspectiva personal.

El primero de ellos es el Dr. Melchor Sánchez Mendiola, actual Coordinador de la Coordinación de Universidad Abierta Innovación Educativa y Educación a Distancia (CUAIEED) de la UNAM, quién no sólo me dio la oportunidad de incorporarme a su equipo de trabajo, donde continuamente existe el desafío en la búsqueda de innovación y reinvención diaria, sino además me impulsó a sacar adelante este asunto pendiente.

La segunda persona a quién dedico estas líneas de agradecimiento es al Dr. Adrián Martínez González, Director de Evaluación Educativa de la CUAIEED, quién me ha transmitido y contagiado su gusto por la investigación, y del que he recibido un incondicional apoyo para iniciarme en este camino tan emocionante y gratificante.

La tercera, la Dra. Silvia Ruiz Velasco Acosta, universitaria con una mente privilegiada, quién amablemente estuvo dispuesta a asesorarme en esta investigación, y me ayudó a enfocar de forma certera en este trabajo.

Asimismo, quiero aprovechar este espacio para agradecerle a la Universidad Nacional Autónoma de México (UNAM) todo lo que me ha dado; recuerdo que uno de mis profesores que mayor huella ha dejado en mí, el Actuario José Enrique Peña, nos decía: "*yo estoy en deuda con la UNAM y quiero devolverle algo de lo mucho que me ha dado*", y ahora que ha pasado el tiempo, en mí ha crecido un sentimiento similar, yo también me siento en deuda con mi amada UNAM, nuestra máxima casa de estudios, por la que desde el fondo de mi ser exclamo un *goya* y por quien me esfuerzo para contribuir en que siga siendo la *Universidad de la Nación*.

Índice general

1. Introducción **3**

2. Pruebas de hipótesis y grandes muestras **4**
2.1. Pruebas de Hipótesis 5
2.1.1. Definiciones 5
2.1.2. Resultados importantes 8
2.1.3. Ejemplo: pruebas de hipótesis con muestras provenientes de una distribución normal 15
2.2. Efecto del uso de grandes muestras en las pruebas de hipótesis 22
2.2.1. Tamaño del efecto 24

3. Meta-análisis **27**
3.1. Modelo de efectos fijos 29
3.2. Modelo de efectos aleatorios 30
3.3. Homogeneidad 33
3.4. Gráfico prototipo de meta-análisis 34
3.5. Relevancia del meta-análisis en esta investigación 35

4. Grandes muestras en el proceso de selección de la UNAM **36**
4.1. Contexto de la problemática con grandes muestras 37
4.2. Material y método 40
4.3. Resultados 40
4.3.1. Afectación a la estimación del efecto 41
4.3.2. Afectación en el dictamen de las pruebas 42

5. Conclusiones **51**

Anexos **54**

A. Anexo 1. Meta-análisis con los grupos de muestra definidos **55**

B. Anexo 2. Códigos en R **86**

B.1. Código para generar el arreglo para realizar los meta-análisis 86

B.2. Código para realizar los gráficos de meta-análisis 90

Capítulo 1

Introducción

Los avances de las tecnologías en tiempos recientes han hecho posible que se puedan realizar labores de investigación con el empleo de grandes volúmenes de información, las organizaciones de todos los sectores ven en el llamado *big data* una fuente de hallazgos que les permita alcanzar beneficios como obtener ventajas competitivas en el mercado en el que se desenvuelven o desarrollar modelos de predicción para entender mejor a poblaciones de interés por citar sólo algunos ejemplos. Sin embargo, esta aparente ventaja que representa la explotación de grandes bancos de datos, ha puesto a prueba las metodologías habituales de inferencia estadística, en especial los contrastes de hipótesis las cuales al ser sometidas a muestras de gran tamaño arrojan resultados que pueden llegar a ser cuestionables. Esta problemática despertó el interés por desarrollar esta investigación donde el objetivo central es determinar un punto de corte en el tamaño de muestra que permita realizar pruebas de hipótesis confiables a través de la metodología del meta-análisis, una técnica que cada vez cobra mayor importancia en el análisis de trabajos de investigación.

Este trabajo consta de cuatro capítulos, en el primero se da un marco teórico a las pruebas de hipótesis y se comienza a plantear la problemática que origina realizar contrastes con grandes tamaños de muestra; en el segundo se proporciona un panorama general sobre la metodología del meta-análisis; en el tercero se aplica la metodologá de meta-análisis a información de un proceso de la UNAM que involucra grandes muestras para intentar contestar la pregunta que de forma implícita se encuentra en el objetivo de esta investigación; en el cuarto capítulo se analizan los resultados y finalmente se presentan las conclusiones de este trabajo.

Capítulo 2

Pruebas de hipótesis y grandes muestras

Es común enfrentarse a situaciones donde se desea conocer el valor poblacional de un parámetro, o se desea tener evidencia acerca de dónde se encuentra. Para ello, se recurre a la extracción de una muestra y con los valores de ésta se estima el valor del parámetro, o se apoya la sospecha sobre su magnitud. En el primer caso se realiza una estimación y en el segundo una prueba o contraste de hipótesis, los cuales son los principales pilares en los que se apoya la inferencia estadística. Hoy en día, son cada vez más las disciplinas que sustentan sus resultados en la inferencia, en especial, con pruebas de hipótesis,un ejemplo de ello se observa en el trabajo de Navarro et. al. [21] donde se realiza un estudio bibliométrico a una muestra de 309 artículos originales de revistas odontológicas indexadas a la red Scielo (Scientific Electronic Library Online) publicados entre 2013 y 2014 y extraídos de un universo de 4,262. Los resultados resaltan que en el 72 % de los trabajos se emplearon métodos paramétricos donde los de mayor frecuencia fueron pruebas post-hoc (36 %); el modelo lineal (27 %) y pruebas t para muestras independientes (9 %). Lo anterior refleja el papel central que juegan las pruebas de hipótesis para dar sustento o rechazar planteamientos teóricos, por lo que se vuelve imperativo no sólo prestar atención a los resultados de las pruebas sino también al planteamiento de éstas y al contexto en el que se están desarrollando.

Por la importancia que representan las pruebas de hipótesis, en este capítulo se retomarán aspectos de este tema con la intención de sentar las bases de este trabajo.

2.1. Pruebas de Hipótesis

Como ya se mencionó, las pruebas de hipótesis son uno de los pilares de la inferencia estadística cuyo uso se generaliza en otras disciplinas para sustentar resultados de investigación. El planteamiento de hipótesis es una de las fases iniciales del método científico que se deriva de la observación. Una hipótesis es un supuesto acerca de una situación que se observa, la cual se apoyará o no en función de los resultados de la experimentación. Un problema de pruebas de hipótesis parte de que la distribución de una población se encuentra afectada por el valor de un parámetro de interés θ el cual es desconocido. A partir de información previa se plantea una hipótesis de su valor y se contrasta con otro supuesto complementario. Para comprobar qué hipótesis es la adecuada, se extrae una muestra de la población y en función de sus valores se decide qué apoyar. De esta manera, una prueba de hipótesis es un contraste de conjeturas respecto a la distribución de una o más variables aleatorias apoyadas en la información que arroja una muestra aleatoria.

A continuación se presentan definiciones relacionadas con este tópico.

2.1.1. Definiciones

Debido a que el tema de esta investigación gira en torno a las pruebas de hipótesis se comenzará con definir lo que es una hipótesis estadística.

Definición 1 *Una hipótesis estadística es una aseveración o conjetura sobre la distribución de una o más variables aleatorias.*

La definición anterior destaca que la hipótesis estadística es la realización de una conjetura sobre la distribución de una o varias variables aleatorias. Normalmente en un problema de pruebas de hipótesis se realiza una aseveración sobre el valor de un parámetro de interés, el cual, por ser éste quien determina el comportamiento de la población, hace que el ejercicio latente sea realizar un supuesto sobre el comportamiento de una variable aleatoria. Ahora bien, la conjetura sobre el parámetro puede ser que éste tenga un valor fijo o un rango de valores, en el primer caso se realiza una aseveración simple mientras que en el otro una compuesta. De esta manera se tiene la siguiente definición.

Definición 2 *Una hipótesis estadística se dice que es* simple *si especifica la distribución de las variables, si no es así, la hipótesis estadística se dirá que es* compuesta.

A las hipótesis estadísticas se les denotará con la letra H seguido de dos puntos y la aseveración que se desea contrastar. Por ejemplo, si en una hipótesis estadística se asevera que el parámetro θ de cierta variable aleatoria tiene un valor inferior a 10, entonces lo anterior se expresa así:

$$H : \theta < 10$$

Una vez que ya se planteó una hipótesis sobre el comportamiento de la variable aleatoria, el siguiente paso es determinar un criterio o procedimiento para apoyar o no esta conjetura. A esta regla se le conoce como *prueba*.

Definición 3 *La prueba de una hipótesis estadística H es una regla o procedimiento para decidir si se rechaza.*

Autores como Mood et al [20] denotan con Υ a la prueba de una hipótesis estadística.

La prueba de una hipótesis estadística Υ implica delimitar un conjunto de valores dónde la conjetura será rechazada

Definición 4 *La región que contiene los potenciales valores de una muestra donde se rechaza la hipótesis estadística H se le denomina región crítica. A la región crítica se le denota como C_Υ.*

La determinación de una región crítica implica, en consecuencia, también establecer una región de no rechazo, es decir, un conjunto de valores de la muestra que originen que la hipótesis no se rechace.

En los problemas de contraste de hipótesis se acostumbra comparar dos conjeturas complementarias, una de ellas se somete a la regla o procedimiento para decidir si se rechaza y se le denomina *hipótesis nula*, y se denota como H_0 a la otra se le nombra como *hipótesis alternativa* y se denota como H_1. En H_0, $\theta \in \Theta_0$ el cual es un subconjunto del espacio parametral mientras que en H_1, $\theta \in \Theta_0^c$. Al tomar una decisión sobre apoyar o no la hipótesis nula se corre el riesgo de rechazarla cuando es verdadera o no rechazarla cuando es falsa, en ambos casos se comete un error. A continuación se definen estos errores.

Definición 5 *Se dice que se comete el* error tipo I *si se rechaza la hipótesis nula cuando ésta es verdadera.*

Definición 6 *Se dice que se comete el* error tipo II *si no se rechaza la hipótesis nula cuando ésta es falsa.*

En todo proceso que involucre el uso de una muestra, debe tenerse presente que se cometerá un error, por lo que debe buscarse sea lo más pequeño posible. Para determinar el tamaño del error se recurre al cálculo de su probabilidad.

Definición 7 *El tamaño del error tipo I es la probabilidad de cometerlo, es decir, es la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera. A esta probabilidad de cometer el error tipo I se le denota como* α.

Definición 8 *El tamaño del error tipo II es la probabilidad de no rechazar la hipótesis nula cuando ésta es falsa. A esta probabilidad de cometerel error tipo II se le denota como* β.

Si se parte de que θ puede estar definido en un conjunto de valores, la probabilidad de que H_0 se rechace dependerá de su valor. Si H_0 fuere cierta, se esperaría que en valores de θ cercanos al fijado en H_0 (θ_0) la probabilidad de rechazar la hipótesis se acerque a cero, por el contrario, valores de θ alejados de θ_0 acercará la probabilidad de rechazo a 1. Este comportamiento puede resumirse en una función que se define a continuación.

Definición 9 *La* función potencia *de una prueba* Υ *es la probabilidad de que* H_0 *se rechace cuando la distribución de la variable de donde procede la muestra tiene como parámetro* θ. *La función potencia se denota como* $\pi_\Upsilon(\theta)$.

El comportamiento ideal de una función potencia es que su valor se acerque a cero conforme θ se aproxime a θ_0 y se aproxime a uno a medida que se aleje de θ_0.

Para cerrar esta sección se definirá un concepto importante el cual es uno de los más empleados en la realización de pruebas de hipótesis.

Definición 10 *Sea* Υ *una prueba de la hipótesis* $H_0 : \theta \in \Theta_0$, *donde* $\Theta_0 \subset \Theta$ *(espacio parametral). El* tamaño de la prueba Υ *de* H_0 *se define como:*

$$\sup_{\Theta_0 \subset \Theta} \pi_\Upsilon(\theta)$$

Al tamaño de la prueba también se le asocia con el tamaño de la región crítica y se le llega a nombrar como *nivel de significancia*.

2.1.2. Resultados importantes

Hasta el momento se han presentado definiciones que permiten encuadrar un problema de prueba de hipótesis y determinar los elementos que lo componen, el siguiente paso es determinar los criterios para rechazar o no una hipótesis, así como evaluar su funcionamiento. En esta sección se muestran definiciones y teoremas que contribuyen a la dictaminación y evaluación de una prueba de hipótesis. Se comenzará con el caso que se contrasten dos hipótesis simples y posteriormente se procederá a una generalización.

Pruebas simples

Para entender los criterios de dictaminación de una prueba de hipótesis se partirá del supuesto de que se están contrastando dos hipótesis simples. En esta situación, supóngase que se conoce que la distribución de una variable aleatoria se encuentra afectada por el parámetro θ, la hipótesis nula es que $\theta = \theta_0$ y la alternativa que $\theta = \theta_1$. Para determinar qué distribución apoyar se extrae una muestra aleatoria y de acuerdo a sus valores se calcula la probabilidad de que provenga de cada distribución, si la probabilidad de que la muestra provenga de una distribución con parámetro θ_0 es mayor a la que tenga parámetro θ_1, no se rechaza la hipótesis nula, en caso contrario se rechaza. En la siguiente definición se recupera lo expuesto en este ejemplo.

Definición 11 (Prueba de razón de verosimilitud simple) *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria proveniente de alguna variable con función de densidad $f_0(\cdot)$ o $f_1(\cdot)$. Se dice que una prueba Υ de $H_0 : X_i \sim f_0(\cdot)$ contra $H_1 : X_i \sim f_1(\cdot)$ es una prueba de razón de verosimilitudes simple si Υ se define como:*

Rechazar H_0 si $\lambda < k$

No rechazar H_0 si $\lambda > k$

Donde:

$$\lambda = \lambda(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f_0(x_i)}{\prod_{i=1}^{n} f_1(x_i)} = \frac{L_0(x_1, \ldots, x_n)}{L_1(x_1, \ldots, x_n)} \tag{2.1}$$

k es una constante no negativa y $L_j(\cdot)$ es la función de verosimilitud por muestrear de la densidad $f_j(\cdot)$; j=0,1.

La definición anterior hace uso de la función de verosimilitud para estimar la probabilidad de que la muestra provenga de una variable cuyo parámetro θ puede ser cualquiera de los dos valores que se están contrastando, si $L_0(\theta) < L_1(\theta)$ entonces es más probable que la muestra provenga de una variable cuya distribución tiene parámetro θ_1, por lo que la razón de verosimilitudes es menor a 1. Ahora bien, *¿qué tan grande debe ser esa diferencia para rechazar la hipótesis nula?* La respuesta es entonces la constante k, que funciona como un punto de corte.

Anteriormente se mencionó que en toda prueba de hipótesis existe el riesgo de cometer un error, y se introdujo el concepto de función potencia. Supóngase que en una prueba Υ, H_0 fuera correcta, es decir, $\theta = \theta_0$ entonces la probabilidad de cometer el error tipo I, α, es la probabilidad de que la muestra se encuentre en la región crítica C^*, es decir:

$$\alpha = P_\theta((X_1, \ldots, X_n) \in C^*) = \pi_\Upsilon(\theta)$$

Ahora bien, supóngase que H_1 fuera correcta, esto implica que $\theta = \theta_1$. Entonces, la probabilidad de cometer el error tipo II, β, ahora es la probabilidad de que la muestra esté en la región de no rechazo $\bar{C}^*$, esto es:

$$\beta = P_\theta((X_1, \ldots, X_n) \in \bar{C}^*)$$

lo que es lo mismo que:

$$\beta = 1 - P_\theta((X_1, \ldots, X_n) \in C^*)$$

o también:

$$\beta = 1 - \pi_\Upsilon(\theta)$$

lo que significa que:

$$\pi_\Upsilon(\theta) = 1 - \beta$$

A continuación se utilizarán estos resultados en la siguiente definición que servirá de

criterio para evaluar el funcionamiento de una prueba.

Definición 12 *Una prueba Υ^* de $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$ se dice que es la* prueba más potente *de tamaño α $(0 < \alpha < 1)$ si y solo si:*

I $\pi_{\Upsilon^*}(\theta_0) = \alpha$

II $\pi_{\Upsilon^*}(\theta_1) \geq \pi_{\Upsilon}(\theta_1)$ *para cualquier otra prueba Υ para la cual $\pi_{\Upsilon}(\theta_0) \leq \alpha$.*

La definición anterior establece que la prueba más potente es la que tiene el menor error tipo II después de fijar el error tipo I.

A continuación se presenta sin demostración el Teorema de Neyman-Pearson que garantiza obtener la prueba más potente si se satisfacen ciertos criterios.

Teorema 1 (Neyman-Pearson) *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria proveniente de $f(x;\theta)$ donde θ es uno de los valores θ_0 o θ_1. Sean α un valor fijo entre 0 y 1; k^* una contante positiva y C^* un subconjunto de valores posibles de la muestra que satisfagan:*

I $P_{\theta_0}[(X_1, \ldots, X_n) \in C^*] = \alpha$

II
$$\lambda = \frac{\prod_{i=1}^{n} f_0(x_i)}{\prod_{i=1}^{n} f_1(x_i)} = \frac{L_0(\theta_0; x_1, \ldots, x_n)}{L_1(\theta_1; x_1, \ldots, x_n)} = \frac{L_0}{L_1} \leq k^* \tag{2.2}$$

Si $(x_1, \ldots, x_n) \in C^*$

y

$$\lambda > k^* si (x_1, \ldots, x_n) \in \bar{C}^*$$

Entonces la prueba Υ^* correspondiente a la región crítica C^* es una prueba más potente de tamaño α de $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$.

El Teorema de Neyman-Pearson garantiza que si la probabilidad de que una muestra aleatoria caiga en la región crítica es α y si la razón de verosimilitudes es menor a k^* cuando los valores de la muestra se encuentran en la región crítica, entonces la prueba Υ^* es la prueba más potente de $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$ y dicha prueba es de tamaño α.

Pruebas compuestas

Hasta el momento, con la intención de entender el concepto y metodología de pruebas de hipótesis, se han mostrado definiciones y resultados aplicados a pruebas simples, ahora se generalizarán estos conceptos y resultados a pruebas compuestas donde se asume que se cuenta con una muestra aleatoria de una variable con distribución $f(x;\theta)$, donde θ pertenece a un espacio parametral Θ y la pruebaΥ consiste en contrastar

$$H_0 : \theta \in \Theta_0$$

contra

$$H_1 : \theta \in \Theta_1$$

Donde Θ_0 y Θ_1 son subconjuntos de Θ y son mutuamente excluyentes, generalmente suele expresarse a Θ_1 como $\Theta - \Theta_0$.

A continuación se presentan las generalizaciones de las definiciones y resultados utilizados en contrastes de pruebas simples y se adicionan otros resultados importantes.

Definición 13 (Razón de verosimilitud generalizada) *Sea $L(\theta; x_1, \ldots, x_n)$ la función de verosimilitud para la muestra $X_1, \ldots, X_n$ que tiene como función conjunta de densidad $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$ donde $\theta \in \Theta$. La* razón de verosimilitud generalizada*, denotada por λ o λ_n se define como:*

$$\lambda = \lambda_n = \lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta\in\Theta_0} L(\theta; x_1, \ldots, x_n)}{\sup_{\theta\in\Theta} L(\theta; x_1, \ldots, x_n)} \tag{2.3}$$

La razón de verosimilitud generalizada se diferencia de la simple en asumir que la muestra proviene de una variable con distribución afectada por un parámetro proveniente de alguno de los dos subconjuntos de valores del espacio parametral considerados en la prueba que tienen la propiedad de ser mutuamente excluyentes. La razón de verosimilitud generalizada al final se convierte en una razón de verosimilitud simple con los valores supremos de cada subconjunto.

Así como se ha mostrado una generalización de la razón de verosimilitud también hay una generalización de la definición de la *prueba más potente*, la cual se presenta a continuación.

Definición 14 *Una prueba Υ^* de $H_0 : \theta \in \Theta_0$ contra $H_1 : \theta \in \Theta - \Theta_0$ se define como la* prueba uniforme más potente de tamaño α *si y sólo si:*

I $sup_{\theta \in \Theta_0} \pi_{\Upsilon^*}(\theta_0) = \alpha$

II $\pi_{\Upsilon^*}(\theta) \geq \pi_{\Upsilon}(\theta)$ *para toda* $\theta \in \Theta - \Theta_0$ *y para cualquier prueba* Υ *con tamaño menor o igual a* α.

Debido a que θ forma parte de un espacio parametral Θ de valores posibles, es factible pensar que éste es un conjunto ordenado, donde podría esperarse que la función potencia tenga un comportamiento monótono de manera que en el valor supremo de Θ_0 se alcance su máximo valor. Ahora bien, al igual que en las pruebas simples, la primera característica de la prueba uniformemente más potente es que en primer lugar se determina el error tipo I, α, y en segundo criterio termina siendo la mejor prueba la que tiene el error tipo II más pequeño.

El tercer resultado importante que se mostró en las pruebas simples fue el teorema de Neyman-Pearson. Para su generalización en las pruebas compuestas es necesario tener presente la distribución de la variable aleatoria de donde proviene la muestra, así como el comportamiento de la razón de verosimilitud. A continuación se mencionan sin demostración los resultados que nos garantizan que de cumplirse ciertas condiciones tanto en la distribución de la variable aleatoria, así como en el comportamiento de la razón de verosimilitud, entonces puede garantizarse realizar la prueba uniforme más potente [20].

Teorema 2 *Sea* $X_1, \ldots, X_n$ *una muestra aleatoria de la densidad* $f(x;\theta)$, $\theta \in \Theta$, *donde* Θ *es algún intervalo.*

Asumiendo que

$$f(x;\theta) = a(\theta)b(x)exp[c(\theta)d(x)]$$

y que existe una función

$$t(x_1, \ldots, x_n) = \sum_{i=1}^{n} d(x_i)$$

I *Si* $c(\theta)$ *es una función monótona creciente en* θ *y si existe un valor* k^* *tal que*

$$P_{\theta_0}[t(X_1, \ldots, X_n) > k^*] = \alpha$$

Entonces la prueba Υ^* *con región crítica*

$$C^* = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) > k^*\}$$

es una prueba uniformemente más potente de tamaño α *de*

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

o de

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

II *Si* $c(\theta)$ *es una función monótona decreciente en* θ *y si existe un valor* k^* *tal que*

$$P_{\theta_0}[t(X_1, \ldots, X_n) < k^*] = \alpha$$

Entonces la prueba Υ^* *con región crítica*

$$C^* = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) < k^*\}$$

es una prueba uniformemente más potente de tamaño α *de*

$$H_0 : \theta \geq \theta_0$$

$$H_1 : \theta < \theta_0$$

o de

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta < \theta_0$$

Definición 15 (Razón de verosimilitud monótona) *Una familia de densidades*

$$\{f(x; \theta) : \theta \in \Theta\}$$

donde Θ *es un intérvalo*
se dice que tiene una razón de verosimilitud monótona *si existe un estadístico*

$$T = t(X_1, \ldots, X_n)$$

tal que la razón

$$\frac{L(\theta^{'}; x_1, \ldots, x_n)}{L(\theta^{''}; x_1, \ldots, x_n)}$$

puede ser una función no creciente de $t(x_1, \ldots, x_n)$ *para cada* $\theta^{'} < \theta^{''}$ *o una función no decreciente de* $t(x_1, \ldots, x_n)$ *para cada* $\theta^{'} < \theta^{''}$.

Teorema 3 *Sea* $X_1, \ldots, X_n$ *una muestra aleatoria de* $f(x; \theta)$ *donde* Θ *es algún intérvalo. Si se asume que la familia de densidades* $\{f(x; \theta) : \theta \in \Theta\}$ *tiene una razón de verosimilitud monótona en el estadístico suficiente* $t(X_1, \ldots, X_n)$ *:*

I *Si la razón de verosimilitud monótona es no decreciente en* $t(x_1, \ldots, x_n)$ *y si existe un valor* k^* *tal que*

$$P_{\theta_0}[t(X_1, \ldots, X_n) < k^*] = \alpha$$

Entonces la prueba correspondiente a la región crítica

$$C^* = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) < k^*\}$$

es una prueba uniformemente más potente de tamaño α *de*

$$H_0 : \theta \geq \theta_0$$

$$H_1 : \theta < \theta_0$$

II *Si la razón de verosimilitud monótona es no creciente en* $t(x_1, \ldots, x_n)$ *y si existe un valor* k^* *tal que*

$$P_{\theta_0}[t(X_1, \ldots, X_n) > k^*] = \alpha$$

Entonces la prueba correspondiente a la región crítica

$$C^* = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) > k^*\}$$

es una prueba uniformemente más potente de tamaño α *de*

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

2.1.3. Ejemplo: pruebas de hipótesis con muestras provenientes de una distribución normal

Resulta ser de uso frecuente en la práctica de la Estadística Inferencial utilizar muestras de una variable que tiene una distribución normal con parámetros μ y σ y se intenta probar si alguno de estos parámetros tiene cierto valor. Esta sección únicamente se enfocará a mostrar como sería una prueba de hipótesis para μ para el caso cuando σ es o no conocido.

Pruebas de hipótesis para μ cuando σ se conoce

Supóngase que de cierta variable con distribución normal se conoce el valor de σ, entonces la función

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Puede expresarse como:

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} e^{\frac{x\mu}{\sigma^2}} e^{\frac{-\mu^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\mu^2}{2\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} e^{\frac{x\mu}{\sigma^2}}$$

donde:

$$a(\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\mu^2}{2\sigma^2}}$$

$$b(x) = e^{\frac{-x^2}{2\sigma^2}}$$

$$c(\mu) = \frac{\mu}{\sigma^2}$$

$$d(x) = x$$

Si se quisiera realizar la prueba:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Dado que la función $c(\mu)$, es monótona creciente en valores de μ mayores a μ_0 y además puede definirse una función t:

$$t(x_1, \ldots, x_n) = \sum_{i=1}^{n} d(x_i) = \sum_{i=1}^{n} x_i$$

por la primera parte del teorema 2, la prueba más potente de tamaño α tiene como región crítica

$$\sum_{i=1}^{n} x_i > k^*$$

donde

$$P_{\mu_0}[\sum_{i=1}^{n} X_i > k^*] = \alpha$$

lo que implica que

$$P_{\mu_0}[\sum_{i=1}^{n} X_i > k^*] = P_{\mu_0}[\frac{\sum_{i=1}^{n} X_i - n\mu_0}{\sqrt{n}\sigma} > \frac{k^* - n\mu_0}{\sqrt{n}\sigma}] = \alpha$$

Es decir

$$P_{\mu_0}[Z > \frac{k^* - n\mu_0}{\sqrt{n}\sigma}] = \alpha$$

Entonces $\frac{k^*-n\mu_0}{\sqrt{n}\sigma}$ es el cuantil $z_{1-\alpha}$ de una distribución normal estandarizada, lo que significa que

$$k^* = n\mu_0 + z_{1-\alpha}\sqrt{n}\sigma$$

De esta manera se rechaza H_0 si

$$\sum_{i=1}^{n} x_i > n\mu_0 + z_{1-\alpha}\sqrt{n}\sigma$$

O bien

$$\bar{x} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

De lo anterior también se deriva como criterio de rechazo

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha} \tag{2.4}$$

Ahora supóngase que se desea realizar la prueba:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

En este caso la función $c(\mu)$, es monótona decreciente en valores de μ menores a μ_0. Al aplicar ahora la segunda parte del teorema 2, la prueba más potente de tamaño α ahora tiene como región crítica

$$\sum_{i=1}^{n} x_i < k^*$$

donde

$$P_{\mu_0}[\sum_{i=1}^{n} X_i < k^*] = \alpha$$

De manera similar a la que se procedió con $H_1 : \mu > \mu_0$ se obtiene como criterio para rechazar H_0 que

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_\alpha \tag{2.5}$$

donde $z_\alpha = -z_{1-\alpha}$

Un tercer caso de prueba que se pudiera realizar es

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Para esta situación se consideran dos escenarios para la hipótesis alternativa: $\mu < \mu_0$ o $\mu > \mu_0$, dado que en uno $c(\mu)$ es una función monótona decreciente y en el otro creciente, la prueba más potente de tamaño α ahora tiene como región crítica

$$\sum_{i=1}^{n} x_i < k^* \cup \sum_{i=1}^{n} x_i > k^*$$

donde

$$P_{\mu_0}[\sum_{i=1}^{n} X_i < k^* \cup \sum_{i=1}^{n} X_i > k^*] = \alpha$$

es decir

$$P_{\mu_0}[\sum_{i=1}^{n} X_i < k^*] + P_{\mu_0}[\sum_{i=1}^{n} X_i > k^*] = \alpha$$

Cada probabilidad se considerara de igual valor, lo que implica que cada una es $\frac{\alpha}{2}$.

Al proceder de manera semejante que las dos pruebas anteriores se rechaza H_0 si

$$\bar{x} < \mu_0 - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

o

$$\bar{x} > \mu_0 + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

lo cual equivale a

$$|\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}| > z_{1-\frac{\alpha}{2}} \tag{2.6}$$

Este resultado implica que la hipótesis nula se acepta si μ_0 con una confiabilidad de $1-\alpha$ está contenido en el intervalo

$$\bar{x} \pm z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

Es importante destacar que el estadístico en (2.4), (2.5) y (2.6) aumenta a medida que n se incrementa, por lo que en muestras grandes la prueba resultará significativa.

Pruebas de hipótesis para μ cuando σ se desconoce

En la sección anterior se mostraron los criterios para realizar pruebas de hipótesis de la media con muestras extraídas de una distribución normal con desviación estándar conocida, en esencia, las tres situaciones vistas confluyen en la comparación del valor estandarizado del promedio muestral respecto al cuantil asociado con la región crítica o de rechazo. Cuando se desconoce la desviación estándar, una alternativa para determinar un criterio de rechazo es el empleo de la razón de verosimilitud generalizada.

Si se parte que la función de verosimilitud de una variable normal con parámetros μ y σ es

$$L(\mu, \sigma; x_1, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

y que los espacios parametrales son:

$$\Theta_0 = \{(\mu, \sigma) : \mu = \mu_0, \sigma > 0\}$$

y

$$\Theta = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$$

Entonces L se maximiza con $\hat{\mu} = \bar{x}$ y $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ que son los estimadores de máxima verosimilitud, al utilizar estos estimadores en L se obtiene:

$$\left(\frac{n}{2\pi \sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

Para el caso de L_0, la función de verosimilitud se maximiza cuando $\hat{\mu} = \mu_0$ y $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{n}$, al sustituir los estimadores en L_0 se llega a la siguiente expresión:

$$\left(\frac{n}{2\pi \sum_{i=1}^{n}(x_i - \mu_0)^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

De esta manera la razón de verosimilitud es:

$$\lambda = \frac{L_0}{L_1} = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}\right)^{\frac{n}{2}}$$

Si en el denominador se realiza lo siguiente:

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu_0)^2$$

Entonces

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$$

Por tanto

$$\lambda = \frac{L_0}{L_1} = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2}\right)^{\frac{n}{2}}$$

lo cual se puede expresar así:

$$\lambda = \frac{L_0}{L_1} = \left(\frac{1}{1 + \frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \right)^{\frac{n}{2}}$$

Por ser λ una función monótona de

$$t^2 = t(x_1, \ldots, x_n) = \frac{n(n-1)\left(\bar{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} = \left(\frac{(\bar{x}-\mu_0)}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n(n-1)}}} \right)^2$$

De esta manera la región crítica $\lambda \leq \lambda_0$ es equivalente a la región crítica $t^2(x_1, \ldots, x_n) \geq k$ [20]. Por lo que la prueba derivada de la razón de verosimilitud es rechazar H_0 si

$$T^2 = \left(\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}} \right)^2 \geq k^2$$

o aceptar H_0 si $-k < T < k$ donde T tiene una distribución t de Student con n-1 grados de libertad [3], por lo que los valores de k serán los que garanticen que la región entre -k y k sea $1-\alpha$ lo que implica que k es el cuantil $t_{1-\frac{\alpha}{2}}(n-1)$.

Del resultado anterior también se deriva un criterio para las pruebas unilaterales empleando como estadístico de prueba T.

Al igual que en el apartado anterior, el estadístico T aumenta su valor conforme n crece, lo que implica que en muestras grandes la prueba será rechazada.

Comparativo de medias de dos poblaciones

Es una situación común enfrentarse a la necesidad de realizar un comparativo del comportamiento de dos poblaciones, por ejemplo, el desempeño de dos grupos de alumnos que estudian cierta carrera donde uno cursa en el sistema escolarizado y el otro en el abierto; o la evolución de pacientes sometidos a diferentes tratamientos médicos, o las ventas de una marca deportiva en clientes que reciben comunicación constante en comparación con los que no las reciben, por citar sólo algunas aplicaciones. Estos casos tienen como interés

común contrastar las distribuciones de dos poblaciones, X_1 y X_2. Supóngase que ambas variables siguen una distribución normal con parámetros μ_1, μ_2, σ_1 y σ_2 y se desea contrastar sus medias por medio de muestras extraídas de cada población de tamaño n_1 y n_2 respectivamente, la hipótesis nula sería

$$H_0 : \mu_1 = \mu_2$$

Y según el contexto del problema la hipótesis alternativa es $H_1 : \mu_1 > \mu_2$ o $H_1 : \mu_1 < \mu_2$ o bien $H_1 : \mu_1 \neq \mu_2$.

La hipótesis nula también equivale probar que $\mu_1 - \mu_2 = 0$. Como se extraen muestras de cada población de tamaño n_1 y n_2 para realizar la inferencia, se emplearán los promedios muestrales como estimadores de las medias. Al tener presente que el promedio muestral se aproxima a una distribución normal con media μ y desviación estándar $\frac{\sigma}{\sqrt{n}}$ conforme la muestra se incrementa y además de que la combinación lineal de n variables normales independientes también tienen una distribución normal con media $\sum_{i=1}^{n} a_i\mu_i$ y desviación estándar $\sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}$ [20], entonces $\bar{X}_1 - \bar{X}_2$ se aproxima a una distribución normal con media $\mu_1 - \mu_2$ y desviación estándar $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, lo que implica que si se asume como cierta H_0 entonces el estadístico

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tendrá una distribución normal estandarizada. Lo anterior se cumple si se conocen los valores de las desviaciones poblacionales, cuando no es así se estiman sus valores con las muestras. Es importante notar que si las varianzas son diferentes se estaría enfrentando una situación conocida como el *problema de Behrens–Fisher* del cual existen propuestas para atacarlo. Una alternativa es utilizar una desviación estándar ponderada de forma que el estadístico de prueba para determinar el criterio de rechazo será:

$$\frac{\sqrt{\frac{n_1 n_2}{n_1+n_2}}(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}}}$$

el cual tiene una ditribución t de student con $n_1 + n_2 - 2$ grados de libertad [20].

2.2. Efecto del uso de grandes muestras en las pruebas de hipótesis

En la sección anterior se abordó el concepto de pruebas de hipótesis, su tratamiento, y el uso de éstas en muestras provenientes de una distribución normal. En este último punto, al derivar los criterios de aceptación, uno de los parámetros que interviene en esta función es el tamaño de la muestra. Hoy en día, gracias a los avances tecnológicos, es factible realizar investigación con volúmenes de información notables de una variable de interés, ejemplo de ello son los trabajos realizados por Martínez-González et al [16] y Campillo-Labrandero et al [2] quienes respectivamente emplearon información de 27,624 y 24,529 alumnos de la Universidad Nacional Autónoma de México (UNAM), otro caso de empleo de cantidades de información notable es la investigación de Callegaro[1] quien utiliza los registros de cuarenta millones de pacientes. El trabajar con grandes muestras por un lado, con base en ley de los grandes números, garantiza tener una mayor precisión en las estimaciones que se realizan pero, por otro lado, de forma contradictoria, para el caso de pruebas de hipótesis, tener una gran cantidad de información no parece ser una ventaja. Para ilustrar esta aseveración supóngase que se tienen dos poblaciones independientes que siguen una distribución normal cada una, para simplificar este ejemplo asúmase que las desviaciones estándares son iguales y conocidas y además se extrae la misma cantidad de elementos de cada población; si se quisiera probar la hipótesis $H_0 : \mu_1 = \mu_2$ se emplearía como estadístico de prueba:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2}{n}}\sigma}$$

Obsérvese que aunque la diferencia de medias muestrales fuera pequeña, el valor del estadístico de prueba se incrementa conforme el tamaño de la muestra aumenta y la hipótesis se rechazará independientemente del valor de μ_1. La figura 2.1 muestra el cambio en el valor del estadístico para una diferencia en medias de una décima y una desviación estándar de 3 unidades. Si la prueba fuera de dos colas para tamaños de muestra superiores a 6,915 elementos la prueba estaría rechazándose por ser mayor el valor del estadístico de prueba que el punto crítico. Valores superiores al punto crítico implican *p-values* cada

vez más diminutos. Si se intentara estimar un intervalo de confianza para el *p-value* al considerar que su valor depende de la muestra [13], debido al volumen de información que se está empleando el rango de valores que incluiría el valor real del *p-value* seguiría siendo minúsculo. En este ejemplo se demuestra que una pequeña variación podría ser estadísticamente significativa si se trabaja con muestras muy grandes.

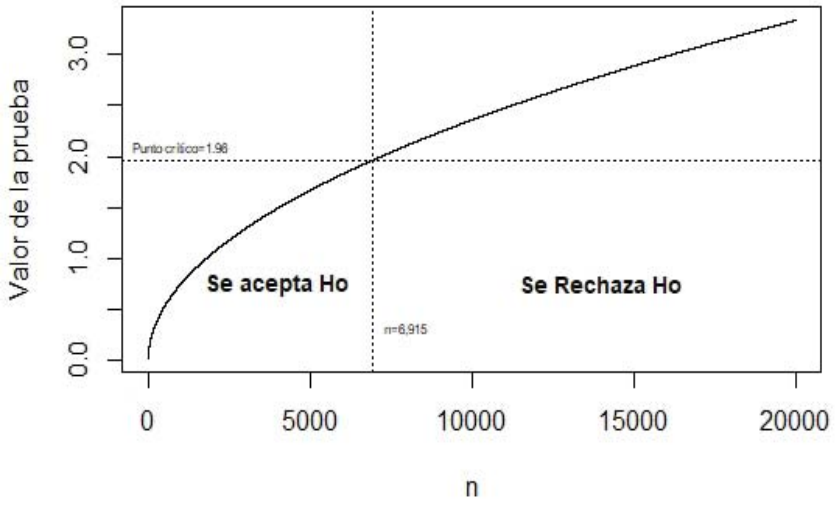

Figura 2.1: Comportamiento del valor del estadístico de prueba para una diferencia de medias muestrales de 0.1 y una desviación estándar de 3 para diferentes tamaños de muestra.

Esta problemática sobre el efecto del uso de muestras muy grandes en las pruebas de hipótesis también es expuesto por Lin [15] en un contexto de investigación en tecnologías, quien destaca, tras realizar una revisión de trabajos de investigación con muestras grandes, que alrededor del 50 % de los estudios justifica sus hallazgos con base en el valor del *p-value* o del signo de los coeficientes del modelo que se ajusta. La revisión también destaca que son pocos los autores que reconocen el problema que acarrea el uso de muestras grandes en el *p-value*, quienes direccionan sus esfuerzos en ajustar los niveles de significancia, recalcular el *p-value* con muestras más pequeñas, enfocarse en los efectos de Cohen en vez de la significancia de la prueba o trabajar intervalos de confianza y auxiliarse con gráficas. Otra alternativa para superar el problema de *p-values* mínimos por el uso de muestras grandes la expone Callegaro [1] quien aplica con éxito la prueba para

diferencias relevantes en un caso real con información de cuarenta millones de pacientes. Para el empleo de esta prueba toma como referencia los trabajos de Cohen [4], Hodges y Lehmann [11], Victor [28] y Wellek [29].

Es un motivo de alarma el hecho de que en la mitad de las investigaciones con grandes muestras se estén justificando hallazgos con base en los valores de *p-values* arrojados por pruebas de hipótesis, incluso como advierte Thompson [25] se corre el riesgo de utilizar el *p-value* como evidencia para la justificación o rechazo de alguna teoría. Este tipo de prácticas ha motivado críticas al uso del *p-value* como las que expone Verdam [27] quien destaca que la prueba de significancia estadística no es necesariamente una evaluación objetiva de los resultados, ya que no muestra el tamaño del efecto, y considera que la significancia de los resultados debiera estar asociada más a un juicio informado que se deriva de la conducción de la investigación y a los hallazgos en vez de los resultados de una prueba. Por otro lado, Marden [18] considera que un *p-value* pequeño no significa necesariamente un rechazo a la hipótesis nula, sino una alerta.

2.2.1. Tamaño del efecto

Una alternativa que complementa a una prueba de hipótesis es estimar el tamaño del efecto. Cohen [6] realiza una propuesta que deja en segundo término la significancia de la prueba, y en vez de ello, sugiere enfocarse en analizar su potencia, ya que ésta, por ser la probabilidad de rechazar H_0 cuando es falsa, se convierte en un criterio para determinar que se va a obtener un resultado que sea estadísticamente significativo. Bajo este enfoque, el éxito de la inferencia dependerá del grado en que la hipótesis nula es falsa, lo que se refleja en el *tamaño del efecto*. De acuerdo a Cohen, el tamaño del efecto es *el grado en que el fenómeno está presente en la población* [26] o bien, *el grado en que la hipótesis nula es falsa* [7] una definición paralela es considerar al tamaño del efecto como *una reflexión cuantitativa de la magnitud de algún fenómeno que se emplea con el fin de abordar una pregunta de interés* [23].

Una prueba de diferencia de medias provenientes de muestras de dos poblaciones de interés puede indicar la existencia o no de una diferencia importante entre dos tratamientos, sin embargo, esta significancia no nos dice lo suficiente sobre el efecto del tratamiento y en ocasiones, en unas áreas más que en otras, se necesita calcular su tamaño[14]. Existen

varias alternativas para estimar el tamaño del efecto, pero una de las más recurridas para examinar la diferencia de medias entre dos grupos independientes es la d de Cohen [26], para su cálculo se utiliza la siguiente fórmula [6] [14].

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_u} \tag{2.7}$$

Donde:
$\bar{X}_1$: promedio de la variable de interés en el primer grupo.
$\bar{X}_2$: promedio de la variable de interés en el segundo grupo.
S_u: desviación unificada.

De acuerdo a Ventura [26] como un marco de referencia, se dice que el efecto es pequeño si $d > 0.20$; mediano si $d > 0.50$ y grande si $d > 0.80$, aunque se tiene el consenso de no considerar estos valores como una regla semejante a la que se sigue para dictaminar una prueba de hipótesis y en vez de ello se sugiere, en caso de cumplirse los supuestos de normalidad, igualdad de varianza y tamaños de muestra equivalentes transformar la d a una medida alternativa como la $U3$, el coeficiente de superposición (OVL), la probabilidad de superioridad (PS) y el número necesario para tratar (NNT) para que el grupo experimental tenga un éxito más (o menos) en comparación al grupo control. Estas medidas se obtienen de la siguiente manera:

$$U3 = \Phi(d)$$

$$OVL = 2\Phi(\frac{-|d|}{2})$$

para grupos iguales

$$PS = \Phi(\frac{d}{\sqrt{2}})$$

y para grupos diferentes

$$PS = \Phi(d\sqrt{\frac{p_1 s_1^2 + p_2 s_2^2}{s_1^2 + s_2^2}})$$

finalmente

$$NNT = \frac{1}{\Phi(d - \Phi^{-1}(CER)) - CER}$$

En las expresiones anteriores Φ es la función acumulada de una Normal estandarizada, d es la medida propuesta por Cohen, p es la proporción de individuos, s la desviación estándar y CER es la tasa de eventos del grupo control.

La d de Cohen también puede transformarse en una correlación punto biserial, o el área bajo la curva normal e incluso una razón de momios como lo muestra el trabajo de Salgado [23].

Ante la problemática de la confiabilidad del resultado de una prueba de hipótesis a consecuencia del uso de grandes muestras, la d de Cohen pareciera ser una alternativa con mayor estabilidad en comparación con estadísticos de prueba afectados directamente por el tamaño de la muestra, por lo que es de interés en esta investigación analizar su comportamiento en diferentes estudios cuando se emplean muestras muy grandes, para ello, el uso del meta-análisis permitirá evaluarlo en campo y compararlo con las pruebas de hipótesis, es por ello, que se destina el siguiente capítulo para hablar del meta-análisis.

Capítulo 3

Meta-análisis

Al comienzo de este trabajo se habló sobre la creciente aplicación de la inferencia estadística en investigaciones de diferentes áreas de conocimiento, se citó como ejemplo el estudio bibliométrico de 309 artículos originales de revistas odontológicas que realizó Navarro et. al. [21]. Otro caso que destaca la importancia del uso de pruebas estadísticas se dio a partir de 1992 en la medicina con un enfoque revolucionario con base en evidencias [10] donde la experiencia clínica debe complementarse con la evidencia documental para favorecer un mejor diagnóstico. Hoy en día, el ritmo de crecimiento de las aportaciones científicas generan una rápida sensación de obsolescencia en las audiencias a las que se dirigen las diversas publicaciones, sin duda, este cúmulo de trabajos es difícil que se pueda consultar en su totalidad, por lo que surge la necesidad de realizar revisiones sistemáticas que permitan reunir en un sólo trabajo los resultados de diferentes investigaciones relacionadas con un tema de interés. Una técnica para realizar esta labor es el *meta-análisis*.

El meta-análisis es una metodología de investigación que consiste en comparar varios estudios relacionados con un tema. Un aspecto, que hay que destacar de esta técnica es que la conformación de la base de estudios para el meta-análisis es un proceso de revisión que puede consumir tiempo considerable, por lo que el planteamiento de la pregunta de investigación, así como la determinación de la variable de interés es fundamental para realizar una búsqueda y revisión más eficiente. Cuando se trabajan con variables cuantitativas, los trabajos que se consideran en el meta-análisis habitualmente comparan dos intervenciones en las que se involucran un grupo *experimental* y otro *control* de los cuales se tiene información de su tamaño de muestra, el promedio y desviación estándar [24]. Como medida del efecto se emplea la diferencia de medias, cuando los estudios utilizan

diferentes escalas de medición se realiza una estandarización de esta diferencia, aunque, en este trabajo, el enfoque será considerar que se trabaja con la misma escala de medición.

Si se considera que los estudios que conforman el meta-análisis se encuentran en la misma escala, entonces para un estudio k, la medida del efecto es

$$\hat{\mu}_k = \hat{\mu}_{ek} - \hat{\mu}_{ck} \tag{3.1}$$

Donde:
$\hat{\mu}_k$: medida del efecto del k-ésimo estudio
$\hat{\mu}_{ek}$: media del grupo experimental del k-ésimo estudio
$\hat{\mu}_{ck}$: media del grupo control del k-ésimo estudio

El efecto en el k-ésimo estudio tiene como varianza

$$Var(\hat{\mu}_k) = \frac{s_{ek}^2}{n_{ek}} + \frac{s_{ck}^2}{n_{ck}} \tag{3.2}$$

Donde:
s_{ek}^2: varianza en el grupo experimental del k-ésimo estudio
s_{ck}^2: varianza en el grupo control del k-ésimo estudio
n_{ek}: tamaño de muestra del grupo experimental del k-ésimo estudio
n_{ck}: tamaño de muestra del grupo control del k-ésimo estudio

Una aproximación a la estimación del efecto en el estudio k con un intervalo de confianza de $1 - \alpha$ $(0 < \alpha < 1)$ es

$$\hat{\mu}_k \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_{ek}^2}{n_{ek}} + \frac{s_{ck}^2}{n_{ck}}} \tag{3.3}$$

Donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $1 - \frac{\alpha}{2}$ de una distribución normal estandarizada.

Una vez que se cuenta por estudio de una estimación de su efecto, el siguiente paso

será estimar uno global, para ello, los modelos que se utilizan para este fin son los de efectos fijos y aleatorios [24] los cuales se comentan a continuación.

3.1. Modelo de efectos fijos

En el modelo de efectos fijos se busca estimar un efecto global bajo el supuesto de que los estudios que componen el meta-análisis provienen de una población homogénea. El modelo en el que descansa este enfoque es el siguiente:

$$\hat{\theta}_k = \theta + \hat{\sigma}_k \epsilon_k \tag{3.4}$$

Donde:
$\hat{\theta}_k$: Estimación del tamaño del efecto en el k-ésimo estudio
θ: Estimación del efecto en la población
$\hat{\sigma}_k$: Estimación de la desviación estándar del efecto del k-ésimo estudio
ϵ_k: Error aleatorio en el k-ésimo estudio el cual es independiente del resto y con distribución normal estándar

El modelo (3.4) indica que el efecto del k-ésimo estudio difiere del global por motivos atribuibles a la muestra, por lo que el promedio de estos efectos se espera coincida con el efecto global.

Si se denota como θ_F al efecto global que se obtiene con el modelo (3.4) el estimador de máxima verosimilitud, y su varianza son [24]:

$$\hat{\theta}_F = \frac{\sum_{k=1}^{K} \frac{\hat{\theta}_k}{\hat{\sigma}_k^2}}{\sum_{k=1}^{K} \frac{1}{\hat{\sigma}_k^2}} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \tag{3.5}$$

$$\hat{Var}(\hat{\theta}_F) = \frac{1}{\sum_{k=1}^{K} w_k} \tag{3.6}$$

Lo que da elementos para calcular un intervalo de confiabilidad $1-\alpha$ $(0 < \alpha < 1)$ para $\hat{\theta}_F$, el cual es

$$\hat{\theta}_F \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{\theta}_F)} \tag{3.7}$$

Como muestra (3.5) el estimador del efecto global es un promedio ponderado de los efectos de los estudios que conforman el meta-análisis, destaca que el peso de cada elemento que conforma la media es la inversa de la varianza del efecto de cada estudio, motivo por lo que a este método se le conoce como el *método de la varianza inversa*. Esta ponderación significa que se otorga mayor peso a los estudios con mayor precisión, lo que se encuentra asociado con el tamaño de la muestra que se utilizó.

3.2. Modelo de efectos aleatorios

El modelo de efectos aleatorios, a diferencia del anterior, supone que las desviaciones observadas entre el efecto de cada estudio en relación con el global no sólo se deben a causas atribuibles al muestreo que se realizó en cada uno, sino a la intervención de otros factores ajenos que se presentan entre ellos, como lo es por ejemplo, la conformación de la muestra, la duración del estudio, o las dosis que se emplearon [22]. En el modelo de efectos fijos se busca estimar el valor de un parámetro, mientras que en el de efectos aleatorios, además de intentar lograr este objetivo, busca estimar el efecto de los diferentes estudios a través de la varianza de alguna distribución, comúnmente la normal.

El modelo bajo este enfoque es :

$$\hat{\theta}_k = \theta + u_k + \hat{\sigma}_k \epsilon_k \tag{3.8}$$

Donde:
$\hat{\theta}_k$: Estimación del tamaño del efecto en el k-ésimo estudio
θ: Estimación del efecto en la población
$\hat{\sigma}_k$: Estimación de la desviación estándar del efecto del k-ésimo estudio
ϵ_k: Error aleatorio en el k-ésimo estudio el cual es independiente del resto y con distribución normal estándar

u_k: Error aleatorio independiente en el k-ésimo estudio con distribución normal de media cero y varianza τ^2

Obsérvese que el modelo de efectos fijos es un caso del modelo de efectos aleatorios cuando $\tau^2 = 0$. Un supuesto que no se debe perder de vista de este modelo es que el error u_k que se observa en nuestra información no está asociado al estudio k, en el sentido de que de volverse a correr el estudio, el error sería una extracción diferente de una distribución $N(0, \tau^2)$.

Para estimar el efecto θ, su varianza y τ^2, la alternativa más empleada es la propuesta por DerSimonian y Laird [24], para ello, se define el siguiente estadístico (de homogeneidad o heterogeneidad):

$$Q = \sum_{k=1}^{K} w_k(\hat{\theta}_k - \hat{\theta}_F)^2 \tag{3.9}$$

Donde:
$w_k = \frac{1}{\hat{\sigma}_k^2}$
$\hat{\theta}_F$ =Estimación del efecto global con el modelo de efectos fijos.

Si $Q < (K-1)$ entonces $\hat{\tau}^2 = 0$ lo que significa que $\hat{\theta}_R = \hat{\theta}_F$

De lo contrario

$$\hat{\tau}^2 = \frac{Q - (K-1)}{S} \tag{3.10}$$

Donde

$$S = \sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k} \tag{3.11}$$

La estimación del efecto aleatorio global es

$$\hat{\theta}_R = \frac{\sum_{k=1}^{K} w_k^* \hat{\theta}_k}{\sum_{k=1}^{K} w_k^*} \tag{3.12}$$

Y la de su varianza

$$\hat{Var}(\hat{\theta}_R) = \frac{1}{\sum_{k=1}^{K} w_k^*} \tag{3.13}$$

Donde $w_k^* = \frac{1}{\hat{\sigma}_k^2 + \hat{\tau}^2}$

Obsérvese que la estimación del efecto global continúa siendo un promedio ponderado de los efectos de los estudios, sólo que el ponderador es el inverso de la varianza aumentada con $\hat{\tau}^2$.

Al igual que en el modelo de efectos fijos es posible construir un intervalo de confianza para $\hat{\theta}_R$

$$\hat{\theta}_R \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{\theta}_R)} \tag{3.14}$$

Debido a la carga de incertidumbre que conlleva la estimación del efecto bajo este modelo, el uso de un intervalo de predicción para futuros estudios que tenga en consideración la varianza entre estudios τ^2 se vuelve necesario. Este rango es

$$\hat{\theta}_R \pm t_{K-2;1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{\theta}_R) + \hat{\tau}^2} \tag{3.15}$$

Donde $t_{K-2;1-\frac{\alpha}{2}}$ es el $1 - \frac{\alpha}{2}$ cuantil de una distribución t de Student con K-2 grados de libertad.

3.3. Homogeneidad

Una vez que se realizaron las estimaciones del efecto con alguno de los modelos, el siguiente paso es determinar en qué grado los resultados pueden generalizarse, es decir, hasta qué punto los resultados de los diferentes estudios pueden combinarse en una medida única, para esto se utilizan medidas de homogeneidad (o heterogeneidad), son de mayor uso las que se muestran a continuación [24].

En párrafos arriba se mostró el estadístico Q en (3.9) el cual puede considerarse como una varianza ponderada de los efectos de los estudios, de manera que si Q tiene un valor grande significa que existe heterogeneidad entre los estudios, es decir existe un τ^2 importante. Para ello, se recurre al resultado de que bajo el supuesto de que $\tau^2 = 0$ entonces Q se distribuye como una Ji-cuadrada con $K - 1$ grados de libertad, por lo que se deriva de ello la siguiente medida

$$H^2 = \frac{Q}{K-1} \tag{3.16}$$

Donde, de cumplirse que τ^2 no es significativa, el valor esperado de H^2 es 1. Si $Q > K - 1$ implica que hay heterogeneidad y mientras mayor sea Q en consecuencia H^2 también lo será. De existir heterogeneidad, es decir, $Q > K - 1$ se utiliza la medida

$$I^2 = \frac{H^2 - 1}{H^2} \tag{3.17}$$

La cual es un escalamiento de H^2, entre mayor sea la heterogeneidad, los valores de I^2 se estarán acercando a 1 y por el contrario, entre más homogéneos sean los resultados de los esudios el valor de I^2 se estará acercando a cero.

3.4. Gráfico prototipo de meta-análisis

Una de las herramientas que le dan valor agregado al meta-análisis son sus gráficas prototipo, éstas permiten visualizar cómo es el comportamiento de los efectos en los estudios considerados, la figura 3.1 ilustra el comportamiento del efecto de dos tratamientos en cincuenta estudios donde se trabajó con dos grupos, experimental y control.

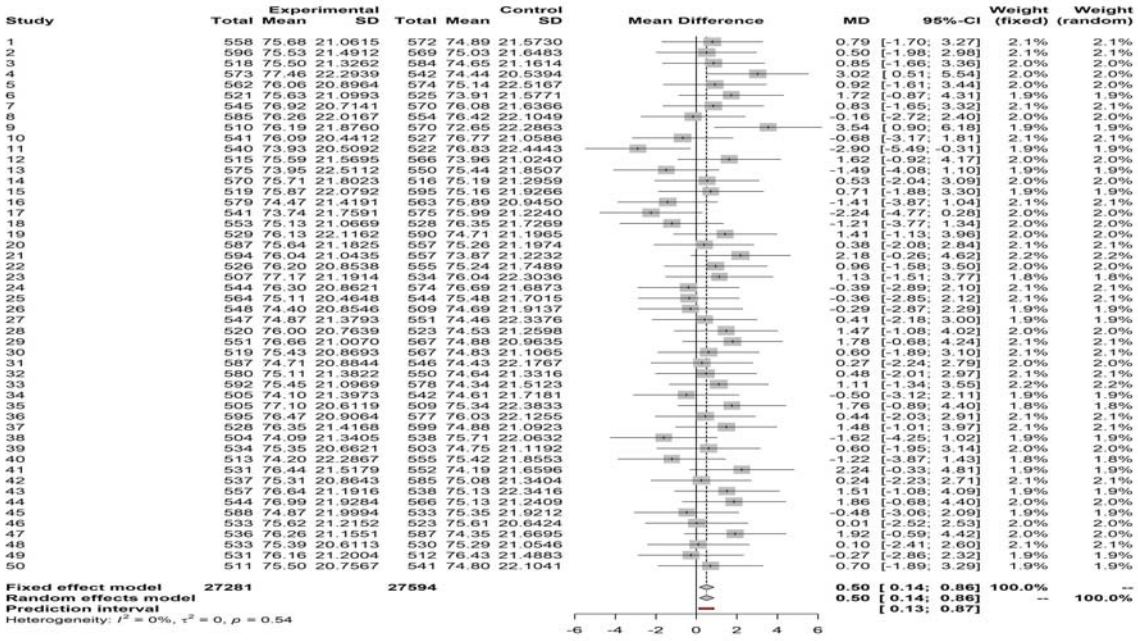

Figura 3.1: Ejemplo de un prototipo de gráfico de meta-análisis.

La figura 3.1 se compone de cinco grandes secciones, la primera la conforma la columna que despliega los estudios que se consideraron en el meta-análisis, en este caso, sólo están enumerados, pero cada campo en esta columna puede contener mayor información del estudio, como *García-Minjares, M. 2020* por ejemplo. La segunda sección tiene información del tamaño de muestra, media y desviación estándar en los grupos experimental y control de cada estudio; la tercera es una representación gráfica de un intervalo de confianza de 95 % para el efecto de cada estudio, entre más a la izquierda se encuentren los intervalos significa que hay un efecto a favor del tratamiento del grupo experimental. En la cuarta sección se muestra la diferencia de medias y los valores extremos del intervalo de confianza de cada estudio. La última parte del gráfico contiene la ponderación de cada investigación

con los modelos de efectos fijos y aleatorios. En la zona inferior de la gráfica se muestra el tamaño total de la muestra y la estimación de la diferencia de medias bajo los modelos de efectos fijos y aletorios así como el cálculo de I^2, $\hat{\tau}^2$ y un intervalo de predicción.

3.5. Relevancia del meta-análisis en esta investigación

El meta-análisis, como ya se mencionó, es una metodología que permite comparar diferentes estudios relacionados con una problemática en común, de forma que se sintetiza información cuantitativa para producir resultados que resuman una investigación. En este capítulo la narrativa se centró en estudios donde se comparan dos tratamientos cuyo efecto θ_k se calcula como la diferencia de medias entre grupos. Para calcular el efecto global se presentaron dos alternativas, en la primera, el modelo de efectos fijos, se asume que los efectos de los estudios individuales difieren del global por cuestiones atribuibles a la muestra, mientras que la segunda opción, el modelo de efectos aleatorios, contempla la variabilidad entre los estudios en la estimación del efecto global.

En esta investigación se tratará de resolver una problemática que engloba comparar dos tratamientos sometidos a poblaciones de tamaño muy grande con el empleo de meta-análisis, donde la estimación del efecto global, fijo o aleatorio, se realizará con el empleo de grupos de muestras de diferentes tamaños que tendrán un rol similar al que jugarían estudios que se llevaran a cabo para analizar el problema y se verá, además, si el método, de aplicarse de forma iterativa, ayuda a identificar un punto de corte donde pruebas realizadas con muestras muy grandes arrojen inferencias válidas. El siguiente capítulo mostrará los resultados de la aplicación del meta-análisis como propuesta de solución a un problema que involucra el empleo de grandes muestras.

Capítulo 4

Grandes muestras en el proceso de selección de la UNAM

En el primer capítulo se hizo mención sobre la problemática que acarrea realizar pruebas de hipótesis con muestras grandes, razón que obliga a tomar con precaución los resultados de estas inferencias. Hoy en día, no existe disciplina donde no sea factible contar con extensos bancos de información que despierten el interés por realizar inferencias en ellos, además de considerarse una fortaleza de investigación utilizar una gran muestra, como lo destacan los trabajos de Campillo-Labrandero et al [2] y Martínez-González et al [16] que analizan el comportamiento del desempeño de generaciones de alumnos de la UNAM, institución que concentra a la fecha más de 360,000 estudiantes activos y 41,332 académicos [9]. Sin duda, el tamaño de población que concentra la UNAM implica el manejo de volúmenes importantes de información en cualquier aspecto que se quiera estudiar de ella, como lo son las historias académicas de sus alumnos, las evaluaciones docentes, los préstamos bibliotecarios e incluso, la población que desea ingresar; este cúmulo de datos permite llevar a cabo diferentes tipos de análisis, la mayor parte de tipo descriptivo, que contribuyan a la toma de decisiones, aunque, no siempre aportan un sustento sólido para rechazar o no un contraste de hipótesis. Debido a la trascendencia de las acciones que se derivan del análisis de la información que se genera en la Universidad, resulta imperativo desarrollar metodologías para mejorar la confiabilidad de las inferencias que se realizan con grandes muestras.

4.1. Contexto de la problemática con grandes muestras

La problemática que se señaló con las pruebas de hipótesis que se realizan con grandes muestras despertó el interés para realizar esta investigación, ya que la participación en diferentes proyectos al interior de la UNAM involucra el análisis de grandes volúmenes de información de varias generaciones de alumnos de bachillerato, licenciatura y posgrado y continuamente nos enfrenta a situaciones donde es necesario establecer si ciertas diferencias de desempeño académico, por citar un ejemplo, debieran ser motivo de preocupación. La problemática en particular que detona esta investigación es el examen de ingreso al nivel medio superior, tema con impacto nacional. En tiempos recientes, cada año, alrededor de 180,000 aspirantes [8] buscan un lugar en alguno de los planteles del bachillerato de la UNAM, lo que la convierte, al menos en este nivel, en la institución educativa con mayor demanda del país, razón suficiente para que este examen de alto impacto concentre la atención de la sociedad. En la aplicación se emplean dos exámenes base equivalentes en diseño y dificultad, y la asignación se realiza en función al número de aciertos que tenga el alumno en el examen que aplicó, dado que los instrumentos son equivalentes, y la población que contesta cada examen tiene características semejantes, entonces es de esperarse que las distribuciones de las calificaciones sean parecidas. En caso de que alguno de los exámenes fuera más fácil, se estaría favoreciendo a la población de alumnos sometidos a esa prueba, lo que podría ser de altas consecuencias para la institución, por lo que parte de la evaluación del examen consiste en validar que no existan diferencias significativas entre aplicaciones.

La figura 4.2 muestra la distribución de las puntuaciones de los alumnos en los exámenes que se aplicaron en el concurso de ingreso al nivel medio superior de 2019, se observa que cada instrumento se aplicó prácticamente a la misma cantidad de aspirantes, poco más de 85,000 en cada uno; en promedio, un alumno que aplicó el primer examen contestó de forma correcta entre 75 y 76 reactivos, en cambio uno del segundo tuvo entre 74 y 75 aciertos; la desviación estándar es de 21.3 en el primer examen y 21.6 en el segundo; la proporción de alumnos con más de 100 aciertos es de 13.8 % en el primero y 13.6 % en el segundo y con menos de 40 aciertos 4.8 % en el primero y 5.3 % en el segundo.

Otro aspecto a destacar es que la distribución de aciertos en ambos exámenes es acampanada y a pesar de la semejanza con una distribución normal, las pruebas de normalidad

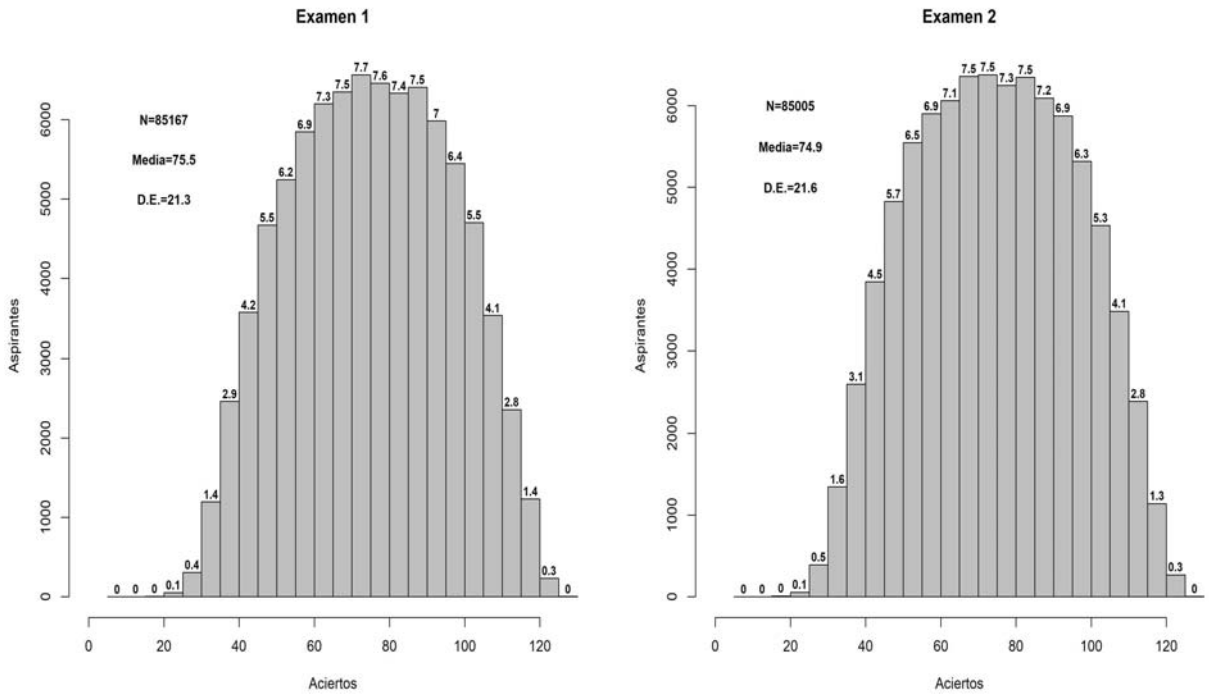

Figura 4.1: Distribución de aciertos de los aspirantes en los exámenes del concurso 2019 para el ingreso al bachillerato de la UNAM. Fuente: elaboración propia con información de la Coordinación de Universidad Abierta Innovación Educativa y Educación a Distancia (CUAIEED) de la UNAM.

rechazan este comportamiento como lo muestra la tabla 4.1.

La tabla 4.1 tiene el resultado de cinco pruebas de normalidad realizadas con el paquete *nortest* de R [12] a las poblaciones que aplicaron cada examen. La prueba de Shapiro-Francia no se pudo realizar por estar definida para muestras entre 5 y 5,000 casos, en el resto, el resultado es un rechazo unánime a la normalidad de las distribuciones al mostrar un *p-value* menor a 0.001. Los resultados del cuadro 4.1 es otro ejemplo del efecto que causa el uso de grandes muestras en pruebas de hipótesis, si se toma como botón de muestra el estadístico Anderson-Darling, éste se obtiene con la fórmula [12]

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1][ln(p_{(i)}) + ln(1-p_{n-i+1})]$$

Prueba	Examen 1	Examen 2
Kolmogorov-Smirnov	D=0.042626; $p < 2.2e^{-16}$	D=0.043771; $p < 2.2e^{-16}$
Anderson-Darling	A=319.19; $p < 2.2e^{-16}$	A=342.54; $p < 2.2e^{-16}$
Cramer-Von Mises	W=47.113; $p = 7.4e^{-10}$	W=50.815; $p = 7.4e^{-10}$
Pearson	P=122480; $p < 2.2e^{-16}$	P=119279; $p < 2.2e^{-16}$
Shapiro-Francia	No definida	No definida

Cálculos realizados con el paquete *nortest* descargado de r-project.org.

Tabla 4.1: Pruebas de normalidad a las poblaciones que aplicaron los exámenes de ingreso al Bachillerato de la UNAM de 2019.

donde $p_{(i)} = \Phi(\frac{[x_{(i)} - \bar{x}]}{s})$

En esta fórmula, Φ es la distribución normal estándar acumulada; $\bar{x}$ y s el promedio y desviación de los datos respectivamente y el *p-value* se calcula del estadístico

$$Z = A(1.0 + 0.75/n + 2.25/n^2)$$

En las fórmulas anteriores, el tamaño de muestra en los denominadores, afecta el resultado a medida que aumenta, por lo que el dictamen de la prueba de normalidad para las poblaciones en estudio no debiera ser concluyente. Por otro lado, si se quisiera sustentar la equivalencia de los resultados mediante una prueba de hipótesis, como ya se ha mencionado, los tamaños de muestra que se emplean, originarán que pequeñas diferencias tengan significancia estadística.

En la búsqueda de un tamaño de muestra lo suficientemente grande que garantice una inferencia aceptable, se propone realizar pruebas de hipótesis con diferentes tamaños de muestra en donde se considerará cada examen como un tratamiento y a las poblaciones respectivas como grupo control y experimental, para ello se aplicará meta-análisis para identificar el momento en que haya una diferencia significativa en el efecto.

El cuestionamiento de investigación de este trabajo es saber si existe la posibilidad

de identificar un punto de corte en el tamaño de la muestra, mediante el meta-análisis, a partir del cual el resultado de una prueba de hipótesis para una diferencia insignificante de dos medias de poblaciones independientes se vuelva significativa. De ser así entonces este punto de corte daría pie para fijar un criterio de evaluación que garantice un resultado que sea estadísticamente válido. En otras palabras, en este trabajo se estaría en la búsqueda de un tamaño esencial de la gran muestra que permitiera realizar pruebas de hipótesis con validez. Algo así como identificar el número de páginas que deberían leerse de un gran libro para conocer lo suficiente sobre su contenido.

4.2. Material y método

Para realizar esta investigación se trabajó con información de los puntajes totales de los dos exámenes que se aplicaron en el concurso de selección 2019 a la educación media superior y se buscó encontrar el punto de corte donde se rechaza la hipótesis nula de que no hay diferencia en el promedio de puntuaciones de las aplicaciones. Los exámenes se consideraron como tratamientos diferentes, para efectos de la rutina utilizada para el análisis, los alumnos que contestaron el primer examen se consideraron como grupo control y los del segundo como grupo experimental, y de ambos se extrajeron muestras. Se definieron 30 rangos de tamaños de muestra, los cuales se presentan en la tabla 4.2, con cada uno se aplicó un meta-análisis con el empleo de los modelos de efectos fijos y aleatorios a 50 pruebas con muestras aleatorias de las poblaciones, cuyo tamaño, determinado de forma aleatoria, se encontró dentro del rango correspondiente y se registró si la prueba fue rechazada. A la par, en cada prueba se calculó la d de Cohen con la intención de comparar su comportamiento con el *p-value* a medida que el tamaño de muestra se incrementa.

4.3. Resultados

Como se mencionó en la sección anterior, se realizaron 30 meta-análisis de 50 pruebas cada uno, en cada ejercicio se utilizaron tamaños de muestra de acuerdo al grupo correspondiente. La figura 4.2 muestra como ejemplo el meta-análisis con el uso de muestras aleatorias entre 1,001 y 2,000 alumnos de cada grupo.

En el meta-análisis con muestras entre 1,001 y 2,000 estudiantes por grupo se observa que la estimación del efecto tanto por el modelo de efectos fijos como aleatorios es la

Grupo	n	Grupo	n	Grupo	n
1	2 - 10	11	901 - 1,000	21	35,001 - 40,000
2	11 - 20	12	1,001 - 2,000	22	40,001 - 45,000
3	21 - 30	13	2,001 - 3,000	23	45,001 - 50,000
4	31 - 50	14	3,001 - 5,000	24	50,001 - 55,000
5	51 - 100	15	5,001 - 10,000	25	55,001 - 60,000
6	101 - 500	16	10,001 - 15,000	26	60,001 - 65,000
7	501 - 600	17	15,001 - 20,000	27	65,001 - 70,000
8	601 - 700	18	20,001 - 25,000	28	70,001 - 75,000
9	701 - 800	19	25,001 - 30,000	29	75,001 - 80,000
10	801 - 900	20	30,001 - 35,000	30	80,001 - 85,000

Tabla 4.2: Grupos de tamaños de muestra.

misma, 0.6 puntos a favor del grupo control, la predicción se estima entre -1.29 a 0.09; una varianza entre estudios, τ^2 de 0.1032 y un nivel de heterogeneidad de 14 %; la muestra total fue de 74,935 para el grupo experimental y 72,509 para el grupo control. Debido a que τ^2 no es significativa se infiere que las muestras que se obtuvieron son similares lo que explica la semejanza del efecto total de ambos modelos.

Al revisar el resto de los meta-análisis (ver anexo A) se valida la homogeneidad de las poblaciones, donde en los ejercicios 1, 12 y 16 se estima una heterogeneidad de 14 % y en el 3 de 18 % mientras que en los 26 restantes fue de cero además de que τ^2 no resulta significativa. Se observa que conforme el tamaño de la muestra aumenta, las estimaciones del efecto favorecen cada vez más al grupo control, lo que a continuación se explica con mayor detalle.

4.3.1. Afectación a la estimación del efecto

La afectación a la estimación del efecto a causa del incremento en el tamaño de muestra se puede apreciar en las figuras 4.3 y 4.4 donde se observa respectivamente el cambio en el intervalo de confianza calculado por el modelo de efectos fijos y aleatorios. Estas gráficas no solo presentan la poca variación entre los resultados de los modelos sino exponen como en muestras menores a 500, las inferencias realizadas sobre las diferencias en las calificaciones promedio de los exámenes no rechazan la hipótesis nula de igualdad de

medias por incluir el intervalo al cero; muestras entre 501 y 10,000 rechazan la hipótesis nula aunque no muestran estabilidad en la estimación del efecto y muestras superiores a 10,000 rechazan la hipótesis nula y presentan estimaciones sin variación importante.

La revisión de los intervalos de predicción de la figura 4.5 muestran un comportamiento parecido excepto que presentan una estimación estable con tamaños de muestra mayores a 15,000.

En la figura 4.6 se presenta el porcentaje de intervalos que contienen al cero en las gráficas de los meta-análisis de la sección A. En la gráfica se observa que con tamaños de muestras mayores a 1,000 (grupo 11) el porcentaje comienza una tendencia a la baja y se aprecia un cambio en la trayectoria cuando se emplean muestras entre 10,001 y 15,000 (grupo 16).

4.3.2. Afectación en el dictamen de las pruebas

El rechazo de la hipótesis nula tiene un comportamiento parecido a una curva logística como se aprecia en la figura 4.7. Hasta el meta-análisis 11 que se realizó con muestras entre 901 y 1,000 estudiantes, en promedio el porcentaje de rechazo fue de 5.09 % el empleo de mayores tamaños de muestra aumenta el rechazo hasta que éste es total en muestras superiores a 50,000. En los grupos 15 y 16 con muestras entre 5,001 y 15,000 se rechaza entre y 40 y 50 % de las pruebas, pero en el grupo 16 que corresponden a muestras entre 10,001 y 15,000 elementos la tendencia de rechazo cambia.

La figura 4.8 presenta la distribución del *p-value* para todos los meta-análisis que se realizaron, la tendencia es que la distribución cambia de un comportamiento uniforme a uno exponencial conforme la muestra se incrementa. A partir de muestras superiores a 1,000 el *p-value* comienza a concentrarse en valores inferiores a 0.05.

Una revisión del comportamiento de la d de Cohen revela tener una tendencia semejante al *p-value* de acuerdo a la figura 4.9. En muestras de 20 o menos, para el problema que se analiza, es factible llegar a dictaminar la existencia de un efecto importante, pero en muestras más grandes la tendencia es concluir que el tamaño de d es cada vez más insignificante conforme se aumenta la muestra.

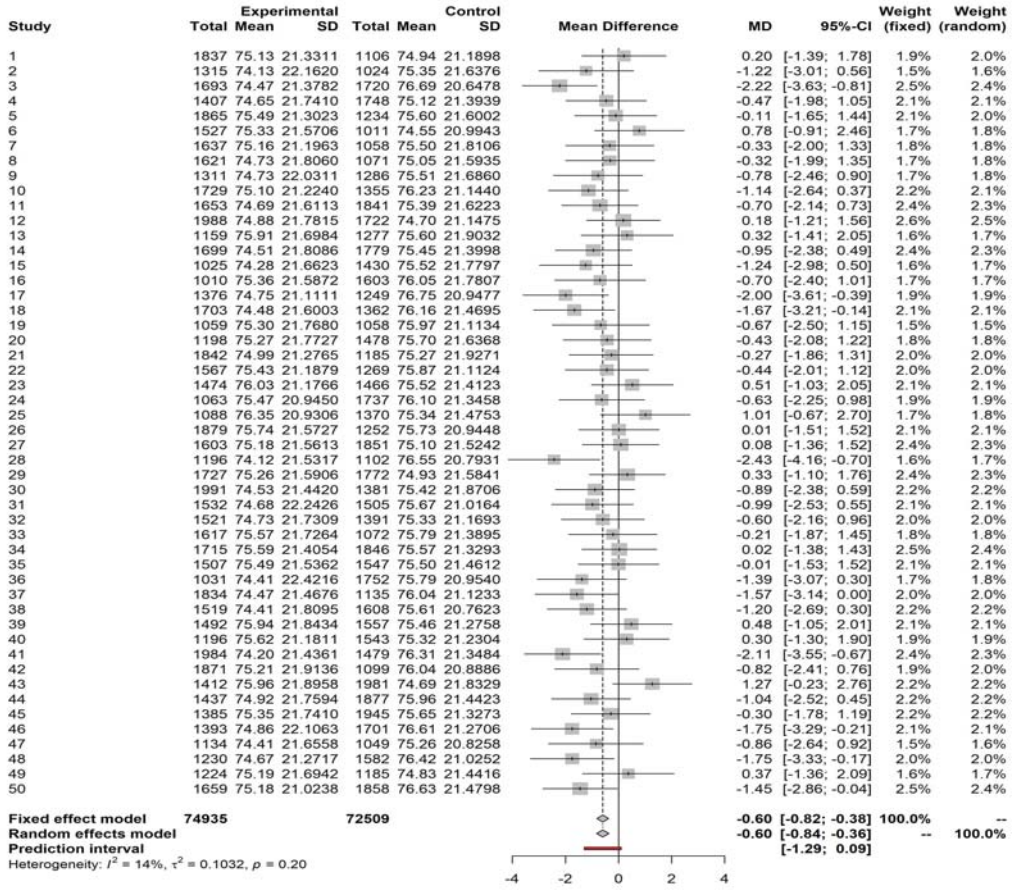

Figura 4.2: Meta-análisis con muestras entre 1,001 y 2,000.

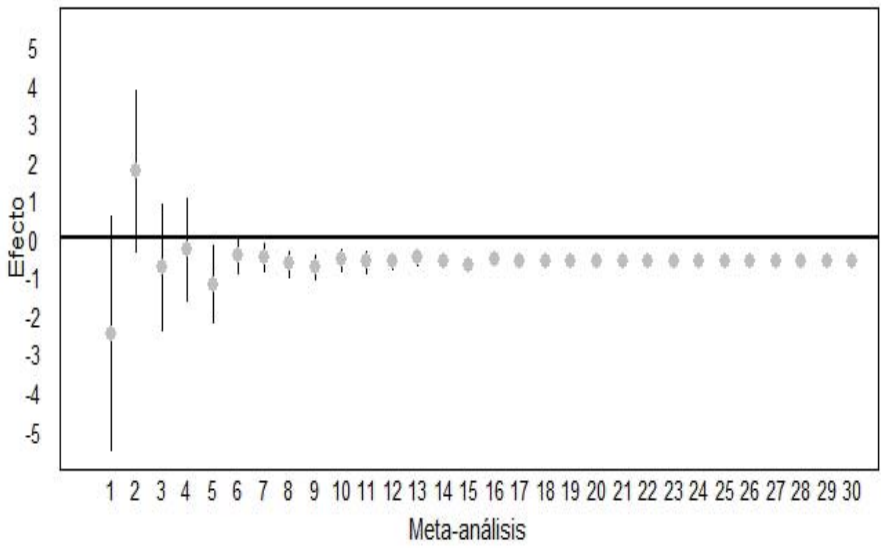

Figura 4.3: Efecto global en cada meta-análisis con el empleo del modelo de efecto fijos.

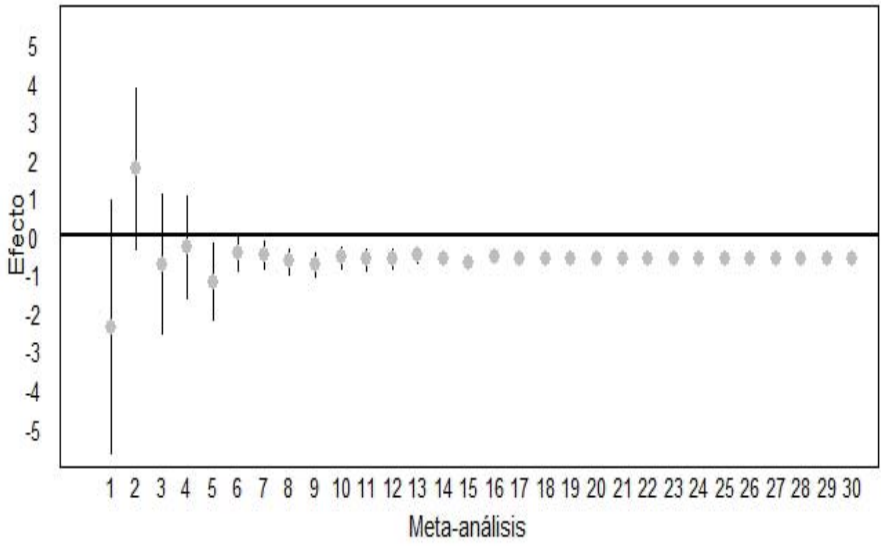

Figura 4.4: Efecto global en cada meta-análisis con el empleo del modelo de efecto aleatorios.

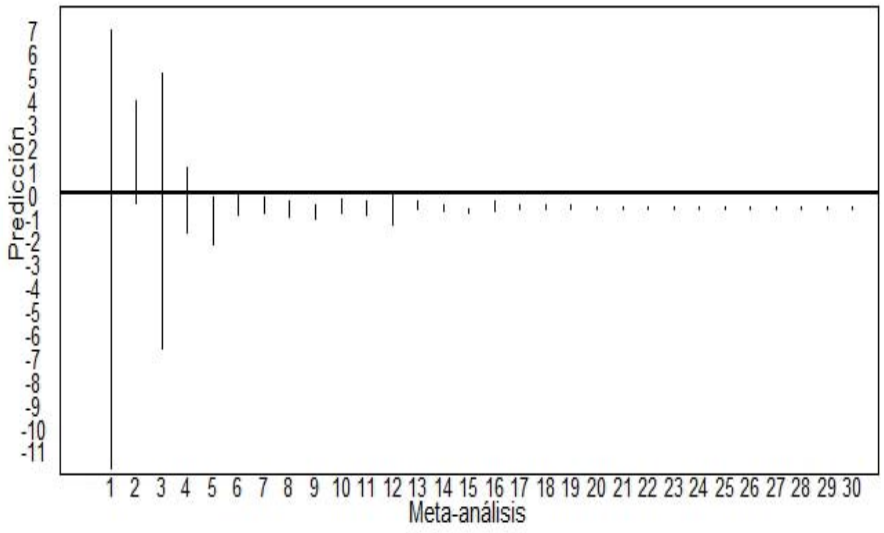

Figura 4.5: Intervalos de predicción en cada meta-análisis.

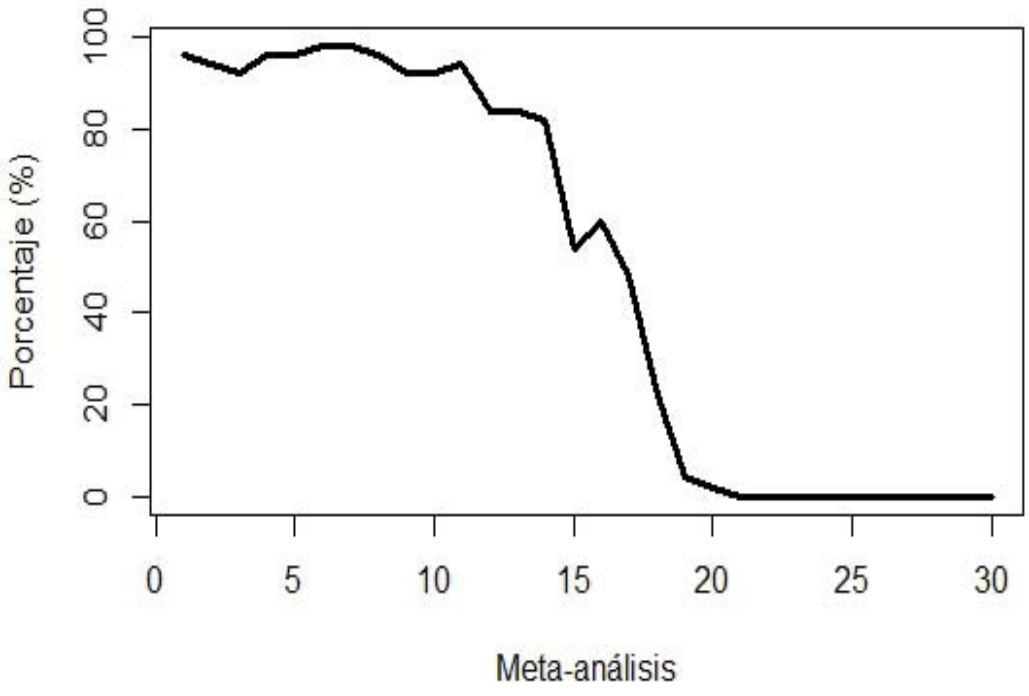

Figura 4.6: Porcentaje de intervalos que incluyen al cero por meta-análisis.

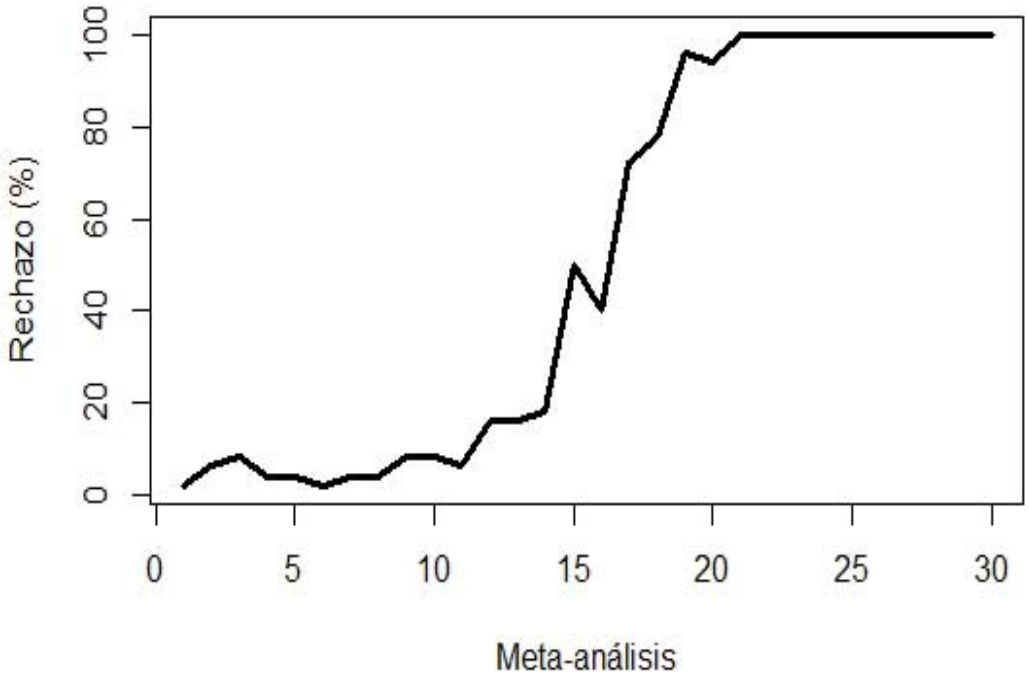

Figura 4.7: Rechazo de la hipótesis nula por meta-análisis.

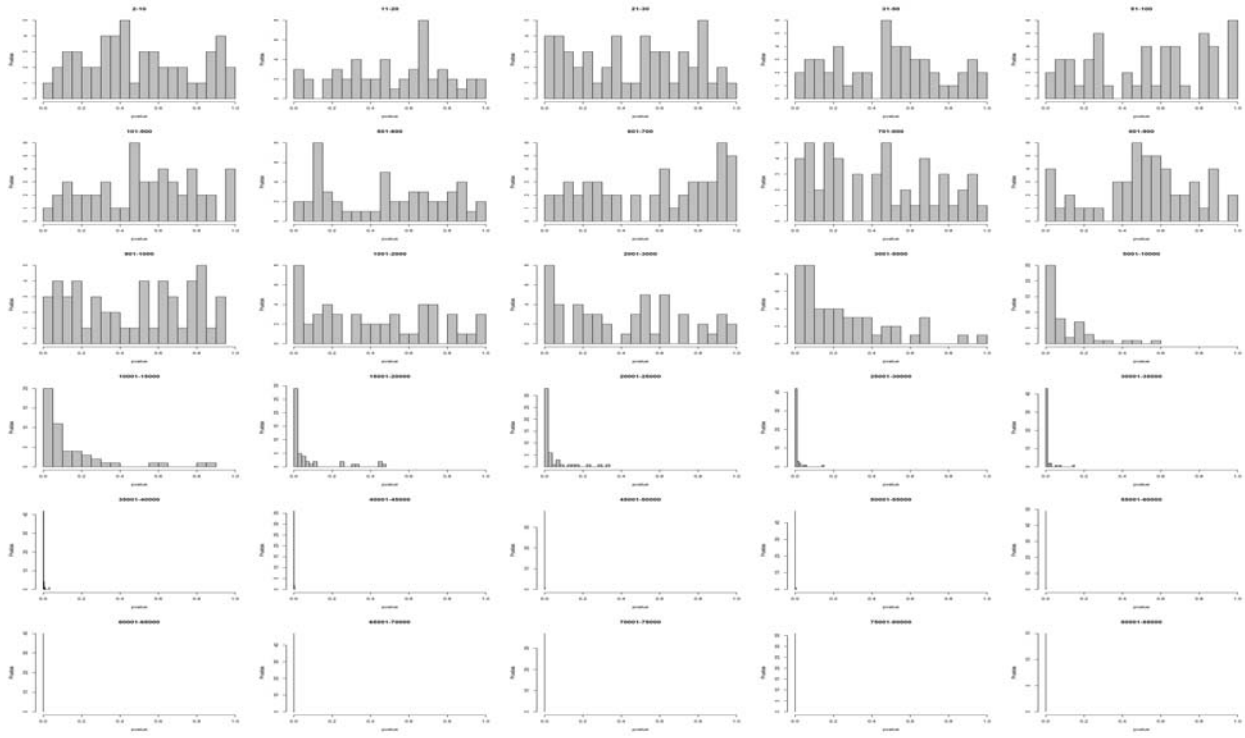

Figura 4.8: Comportamiento del *p-value* para cada grupo de muestras.

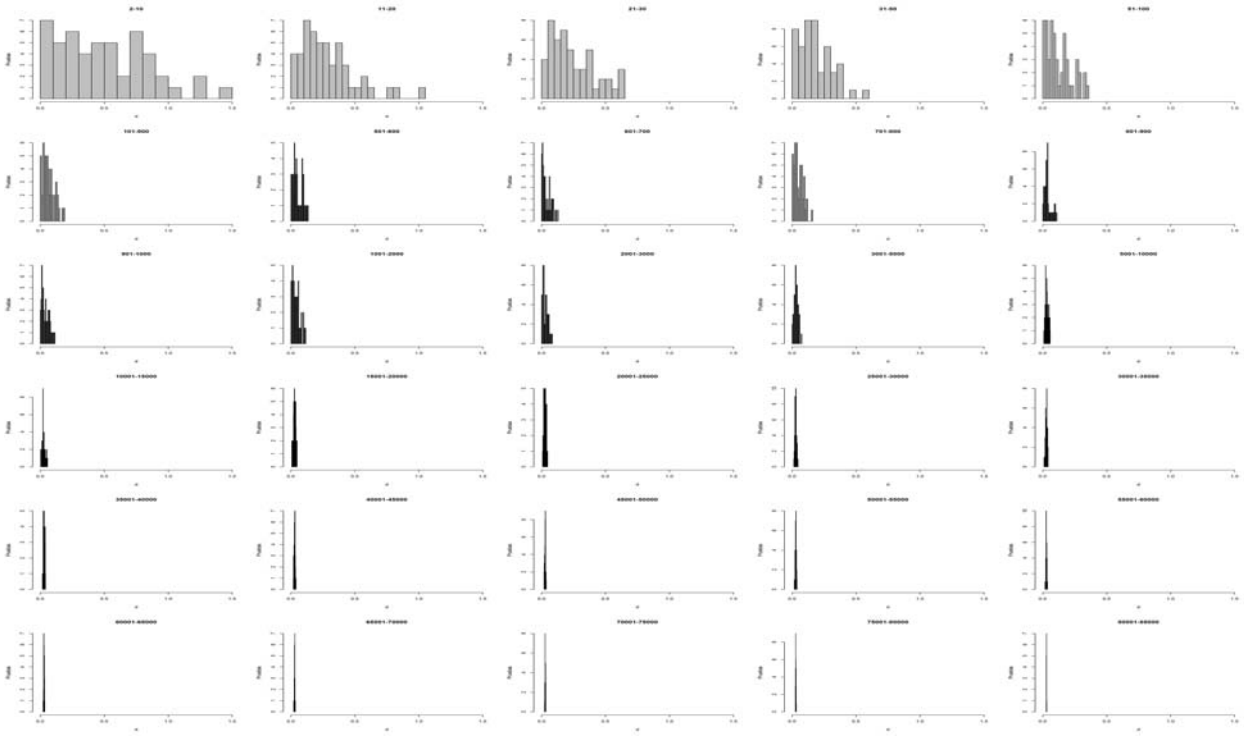

Figura 4.9: Comportamiento de la d de Cohen para cada grupo de muestras.

Capítulo 5

Conclusiones

El empleo de la Estadística Inferencial es un componente central de gran parte de la producción científica en diferentes áreas de conocimiento donde la estimación de parámetros y las pruebas de hipótesis son las principales herramientas para apoyar o no argumentaciones respecto a un tema en particular; la realización de estas tareas implica trabajar con muestras de poblaciones de interés. Hoy en día, a consecuencia de los adelantos en las tecnologías, es factible que el investigador disponga de grandes volúmenes de información relacionadas con una o diferentes variables, este hecho, sin duda representa una fortaleza en trabajos donde se requiere describir los datos o realizar estimaciones, pero cuando se trata de probar hipótesis representa un factor en contra, en la sección 2.2 de este trabajo se mostró que una prueba de igualdad de medias se dictamina como significativa aún cuando éstas tuvieran una diferencia minúscula conforme aumente el tamaño de la muestra y en la sección 4.1 se ilustró como un conjunto de datos que siguen una distribución acampanada, sus pruebas de normalidad resultan significativas como consecuencia también de los tamaños de muestra, aunque no hay que perder de vista que los valores de estas variables se encuentran acotados entre 0 y 128 aciertos, lo que hace que sus colas sean cortas a pesar de ser acampanada.

Una vez que se identificó la problemática que conlleva el uso de grandes muestras en pruebas de hipótesis, en este trabajo se planteó como pregunta de investigación si es posible identificar un punto de corte o de inflexión en el tamaño de muestra a partir del cual se identifique un cambio en el dictamen de las pruebas de hipótesis, para ello se trabajó con datos de los puntajes de los exámenes que se aplicaron a aspirantes a ingresar al bachillerto de la UNAM en 2019, con la intención de identificar ese punto de corte y

por el otro validar la poca diferencia que existe entre los exámenes de admisión. Para intentar responder a la pregunta de investigación se realizaron treinta meta-análisis con cincuenta pruebas donde cada uno se aplicó con un rango de tamaño de muestra distinto y se ajustaron los modelos de efectos fijos y aleatorios.

Los resultados que se obtuvieron de los meta-análisis confirmaron en primer lugar la homogeneidad de las poblaciones ya que las diferencias en las estimaciones realizadas con el modelo de efectos fijos no son muy diferentes a las del modelo de efectos aleatorios, el cual registró en los meta-análisis 1, 12 y 16 una heterogeneidad de 14 % y en el 3 de 18 % mientras que en los 26 restantes fue de cero, por otro lado, la estimación del efecto global de aplicar el examen 1 respecto al 2 estimó el intervalo que incluye el valor real con tamaños de muestra mayores a 10,000 (grupo 16) aunque a partir de muestras mayores a 500 casos por tratamiento (grupo 7) las estimaciones del efecto global ya no incluyen al cero en el intervalo.

La revisión del comportamiento de la tasa de pruebas rechazadas con las muestras que se utilizaron en los meta-análisis demostró que tabajar con tamaños de muestra máximos de 1,000 elementos nos proporcionan una tasa de rechazo acorde a los niveles de significancia que se utilizan con regularidad y una vez que se supera ese límite el rechazo se dispara, observándose un cambio en la tendencia con muestras que superan los 10,000 casos (grupo 16). Un hallazgo que complementa este comportamiento es que la distribución del *p-value* se modifica de una distribución uniforme a una exponencial conforme aumenta el tamaño de la muestra y es después de muestras de 1,000 elementos donde empieza a ocurrir una mayor frecuencia de valores inferiores a 0.05. En cuanto al comportamiento de la d de Cohen, medida que se comentó como alternativa en lugar del *p-value*, este valor se encontró que tiende a ser más diminuto a medida que la muestra crece y el investigador o tomador de decisiones podría interpretar que el tamaño del efecto es insignificante.

Las sugerencias que pueden derivarse de los resultados de esta investigación, al menos para el caso de la revisión de los resultados del examen de admisión al bachillerato de la UNAM, es que si se quisiera, como un primer acercamiento, probar la equivalencia de los exámenes bastaría realizar una prueba con muestras de 1,000 alumnos por examen, la cual proporciona una estimación aceptable del efecto.

Se identifican tres razones que dan relevancia a los resultados de esta investigación, en primer lugar, los hallazgos de este trabajo aportan evidencia para cuestionar la validez de los resultados e interpretaciones que se derivan de pruebas de hipótesis realizadas con muestras grandes, en especial cuando existen diferencias minúsculas entre medias de poblaciones; en segundo lugar, se identificaron tamaños de muestra que además de dar respuesta a la pregunta de investigación, sirven de guía para el diseño de un proceso de evaluación de la aplicación del examen de admisión al bachillerato de la UNAM y finalmente, fija precedentes para futuras líneas de investigación en el tema, en especial el hallazgo del cambio de la distribución del *p-value*. Un rasgo de innovación que se presentó en este trabajo es el uso de la metodología del meta-análisis para la construcción de evidencia para la problemática que se señaló, de lo cual, no se identificaron tratamientos similares en la literatura reciente.

Como cualquier investigación, este trabajo tiene áreas de oportunidad, una de ellas es poder replicar el análisis con más generaciones y probar diferentes tamaños de muestra y de pruebas en cada estudio, así como intentar encontrar un valor puntual de corte. Otro aspecto que no se está considerando, en el caso de las calificaciones del examen de admisión al bachillerato de la UNAM, es que los alumnos aplican en diferentes sedes, días y turnos, por lo que el esquema de muestreo es otro aspecto que podría intervenir en el resultado de una prueba de hipótesis con grandes muestras.

Sin duda, como ya se mencionó, este trabajo deja una contribución que marca pautas para futuras líneas de investigación relacionadas con pruebas de hipótesis para grandes muestras, un tema que en opinión personal se debe continuar trabajando para establecer las bases de una Teoría de Estadística Inferencial para grandes muestras que responda a las exigencias de estos tiempos del *Big Data*.

Anexos

Apéndice A

Anexo 1. Meta-análisis con los grupos de muestra definidos

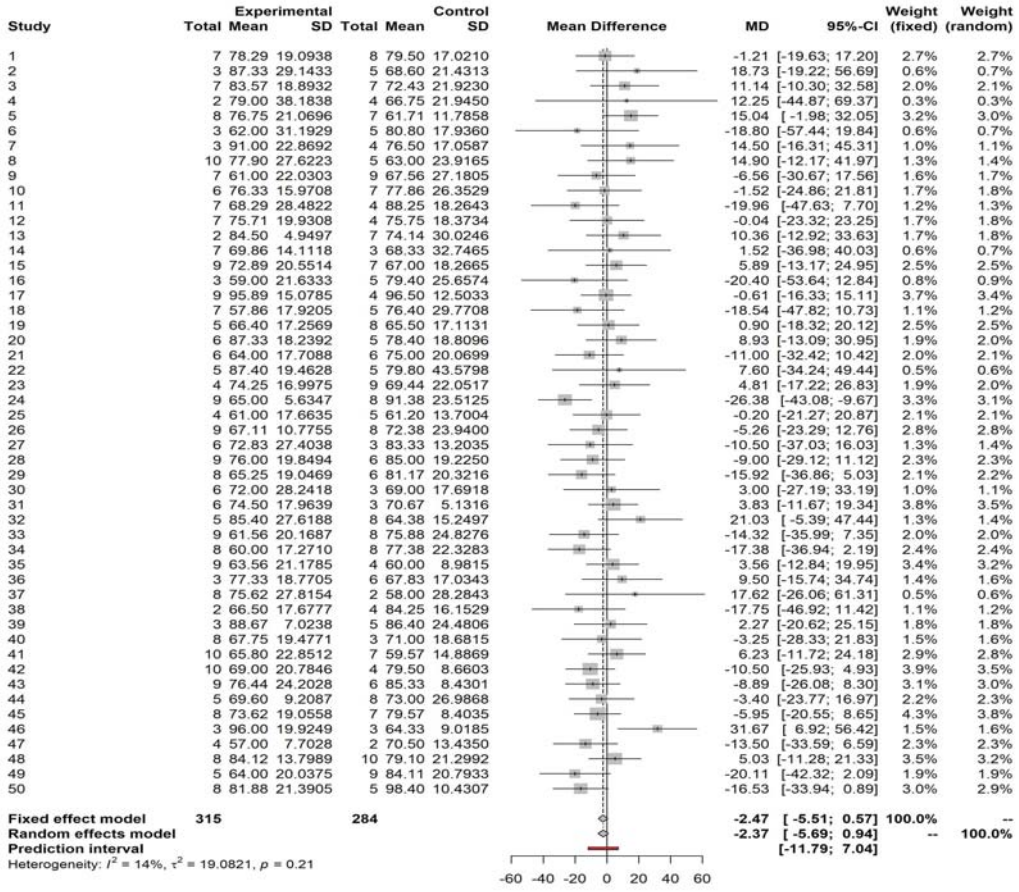

Figura A.1: Meta-análisis con muestras entre 2 y 10.

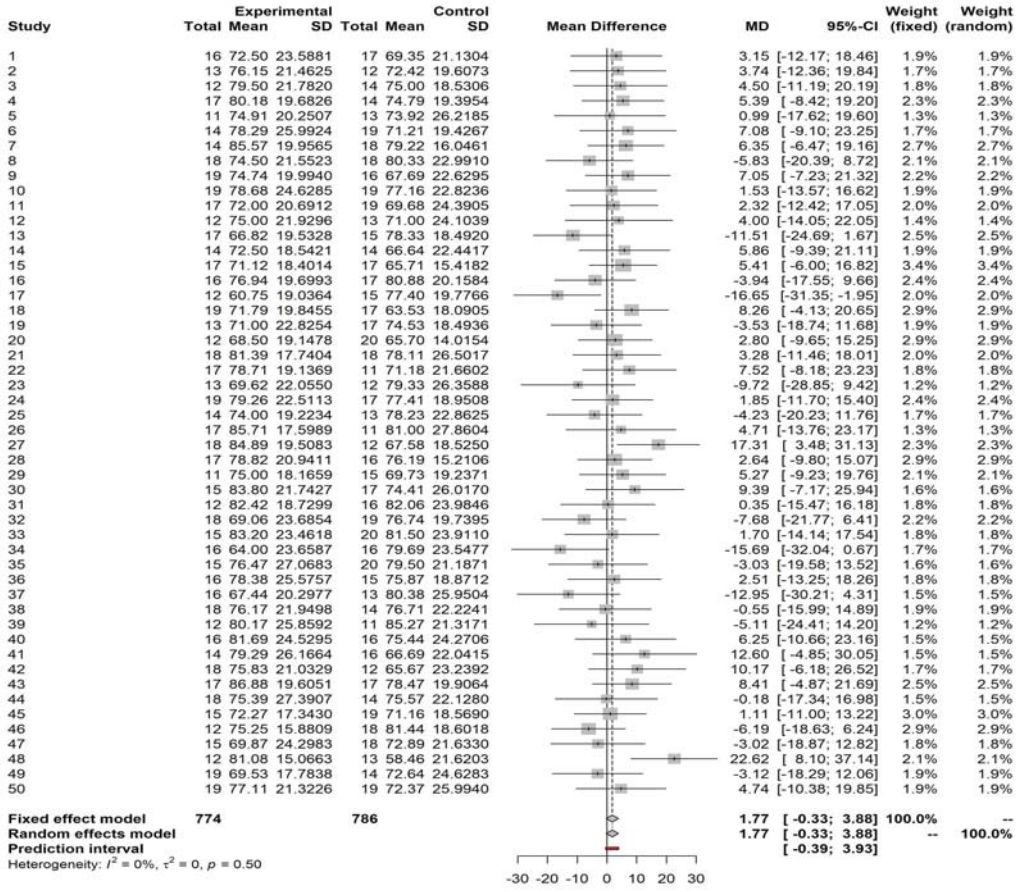

Figura A.2: Meta-análisis con muestras entre 11 y 20.

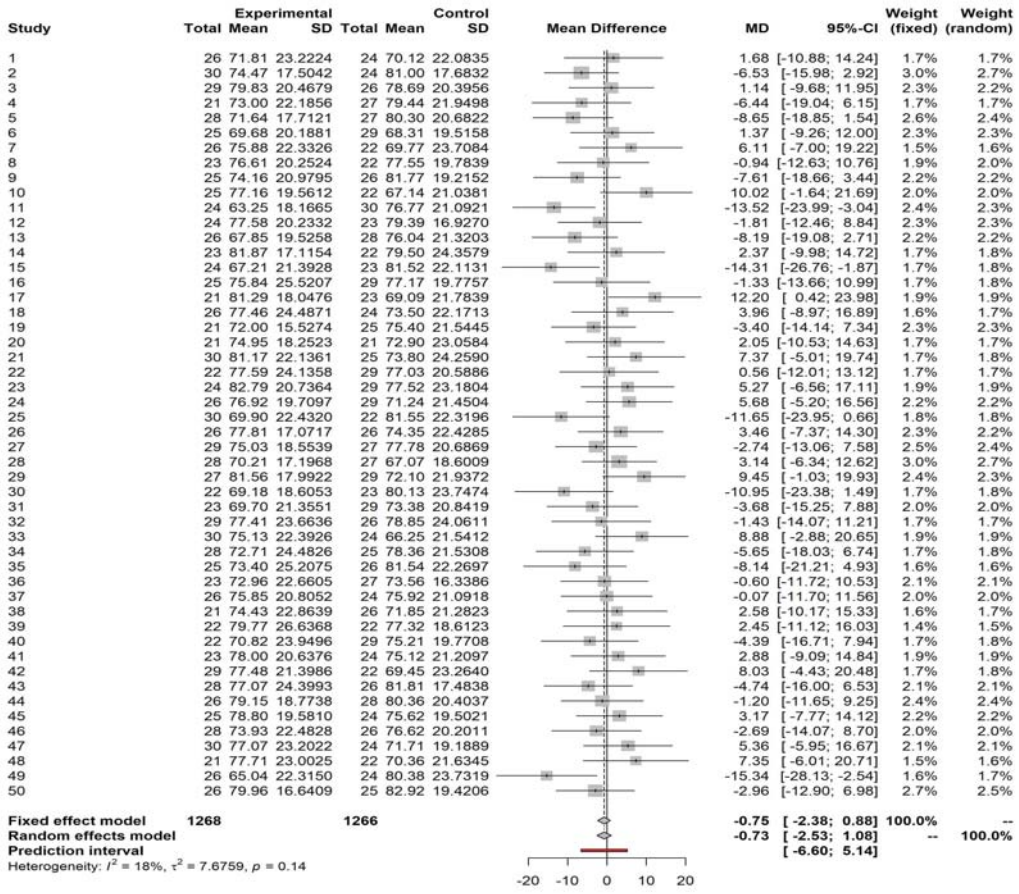

Figura A.3: Meta-análisis con muestras entre 21 y 30.

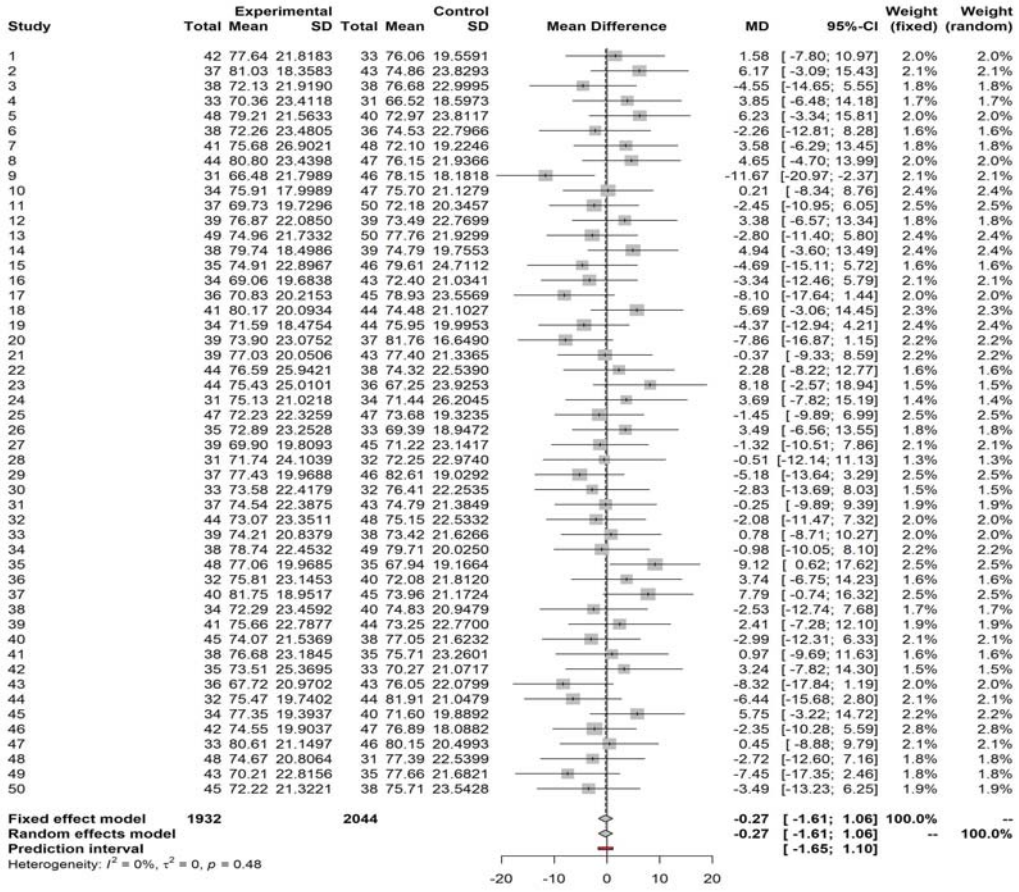

Figura A.4: Meta-análisis con muestras entre 31 y 50.

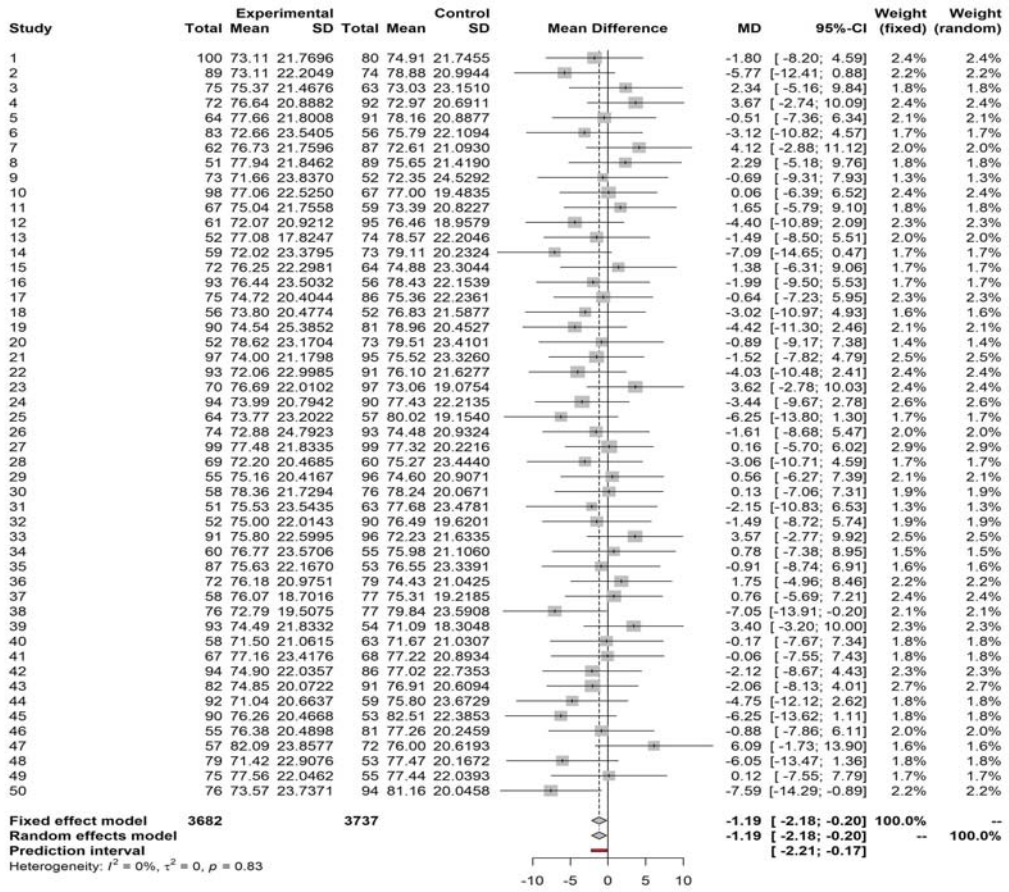

Figura A.5: Meta-análisis con muestras entre 51 y 100.

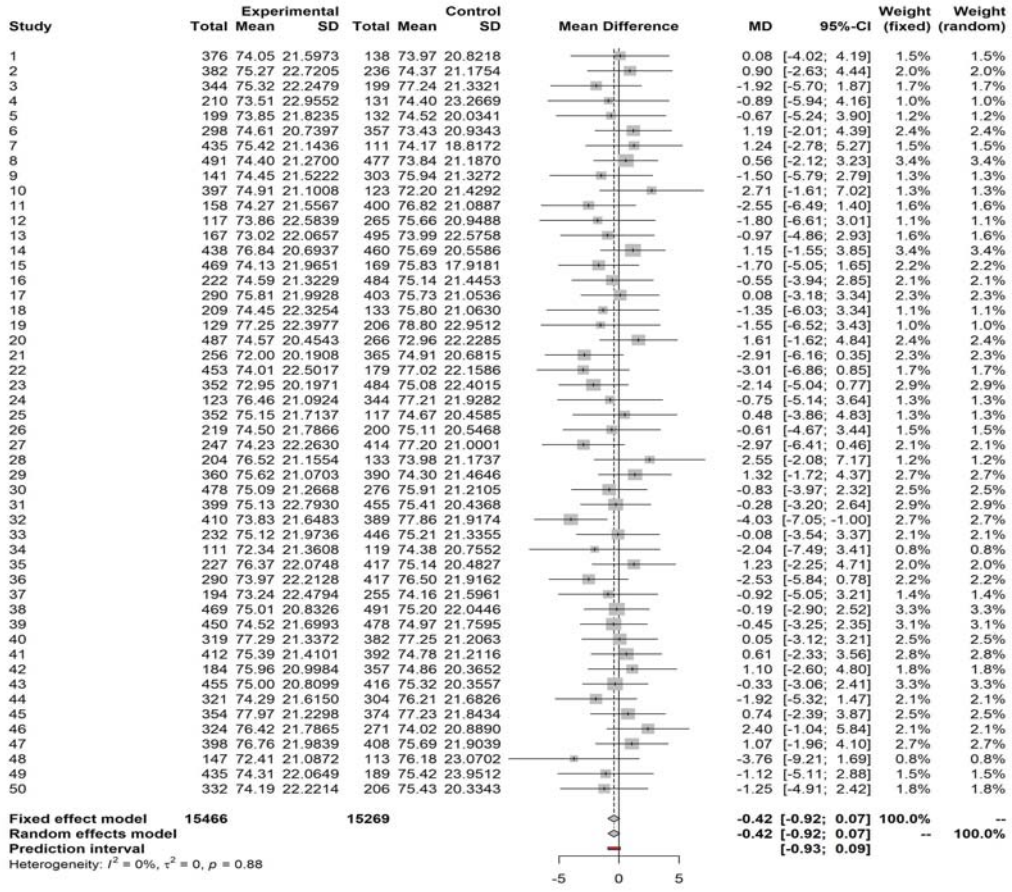

Figura A.6: Meta-análisis con muestras entre 101 y 500.

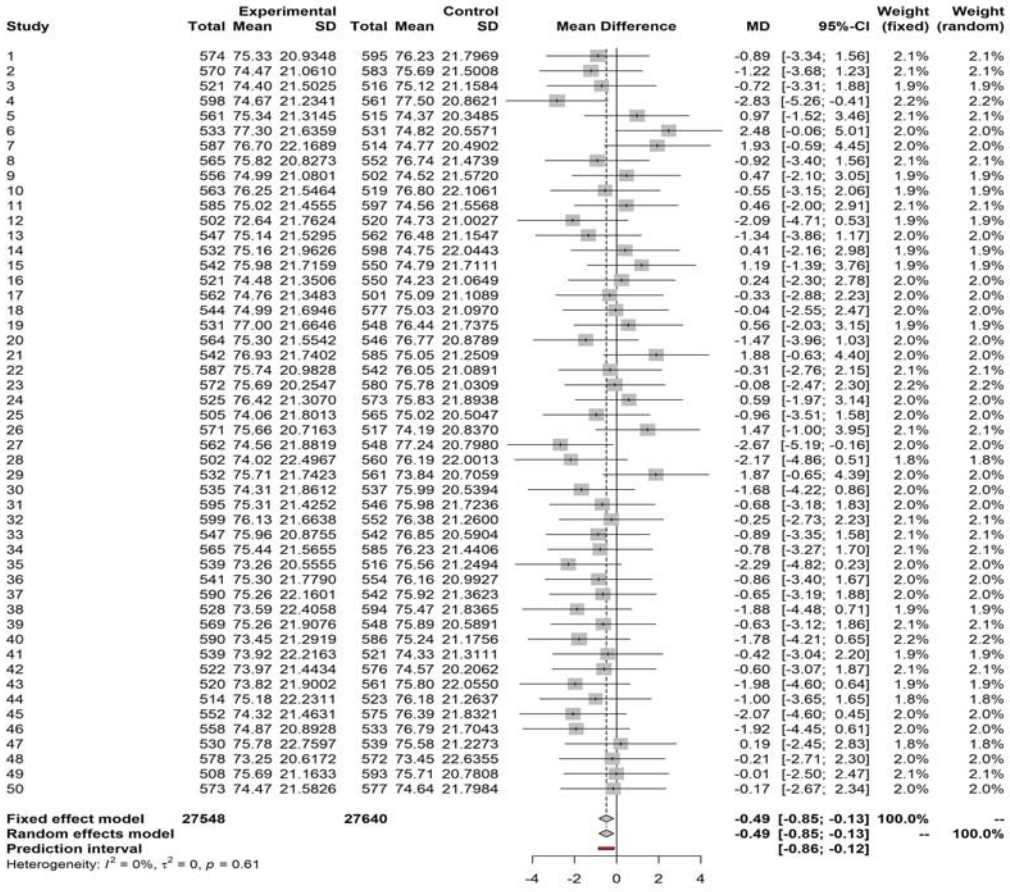

Figura A.7: Meta-análisis con muestras entre 501 y 600.

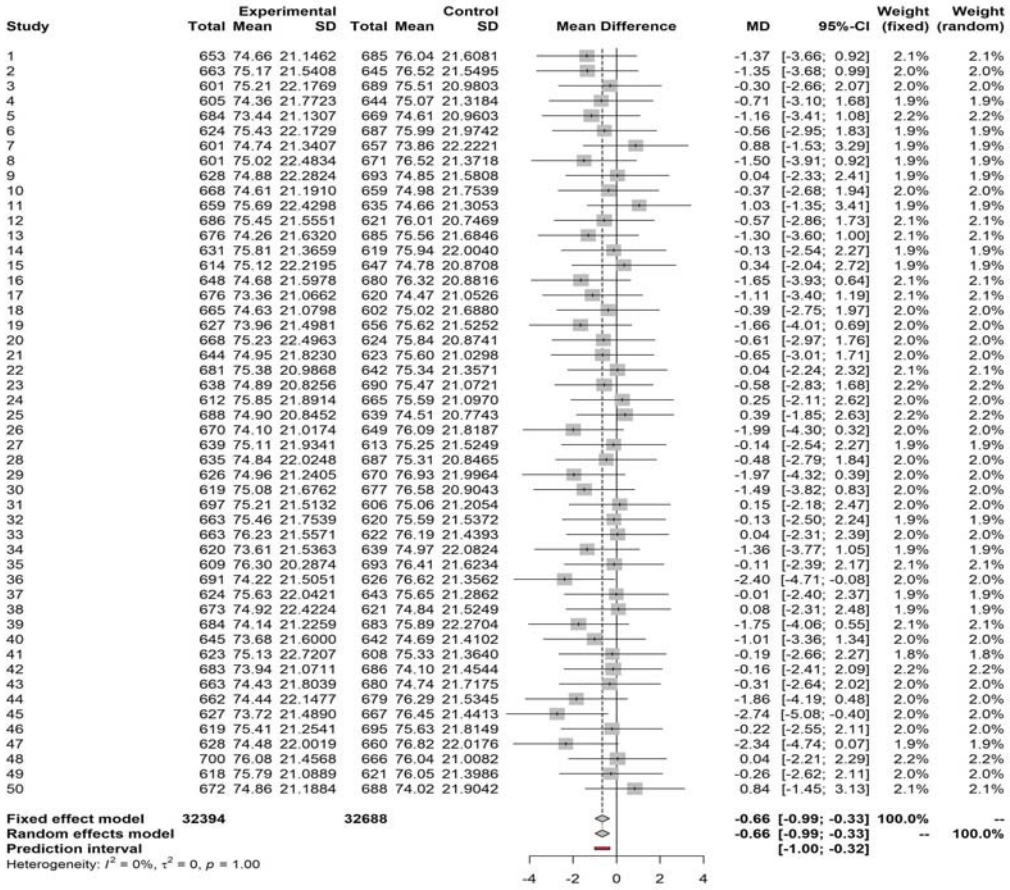

Figura A.8: Meta-análisis con muestras entre 601 y 700.

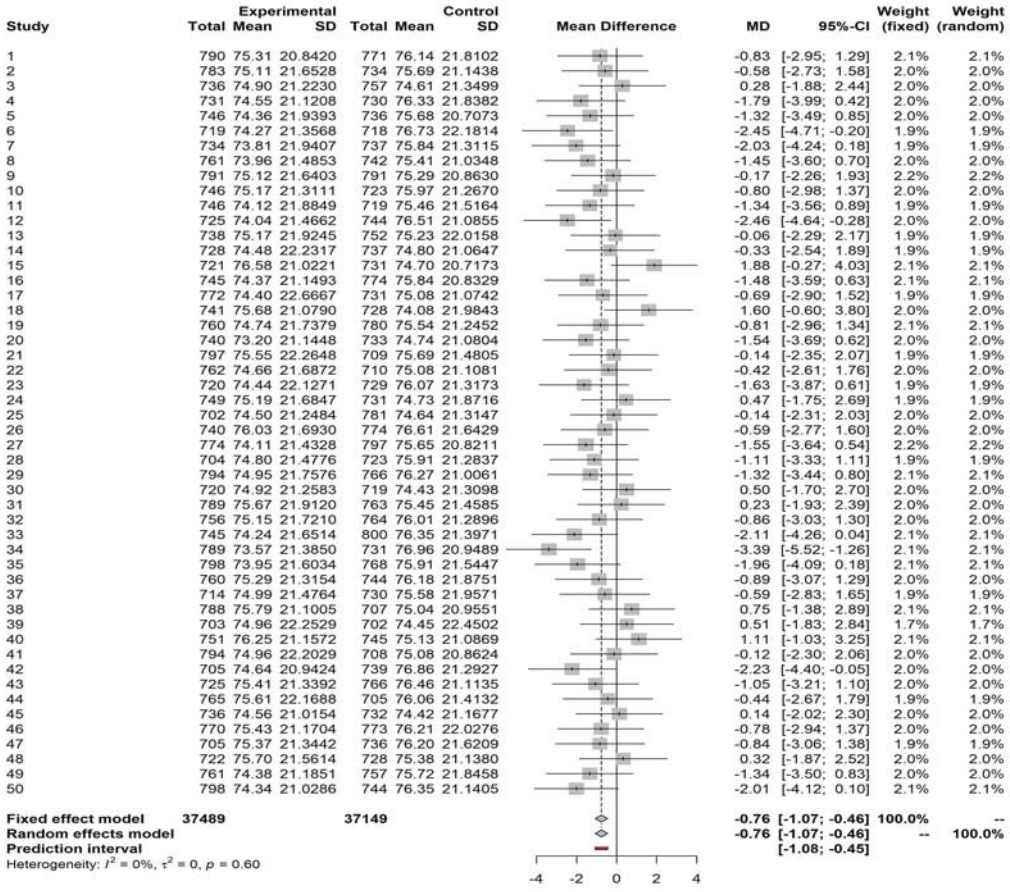

Figura A.9: Meta-análisis con muestras entre 701 y 800.

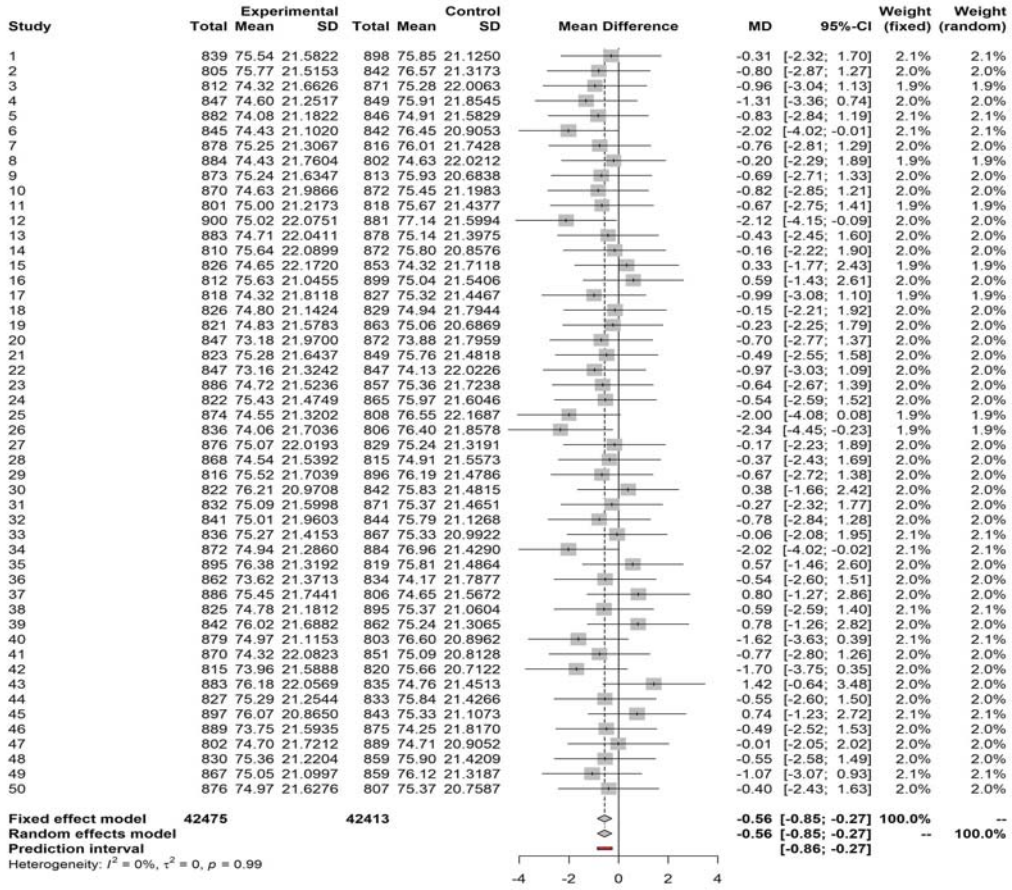

Figura A.10: Meta-análisis con muestras entre 801 y 900.

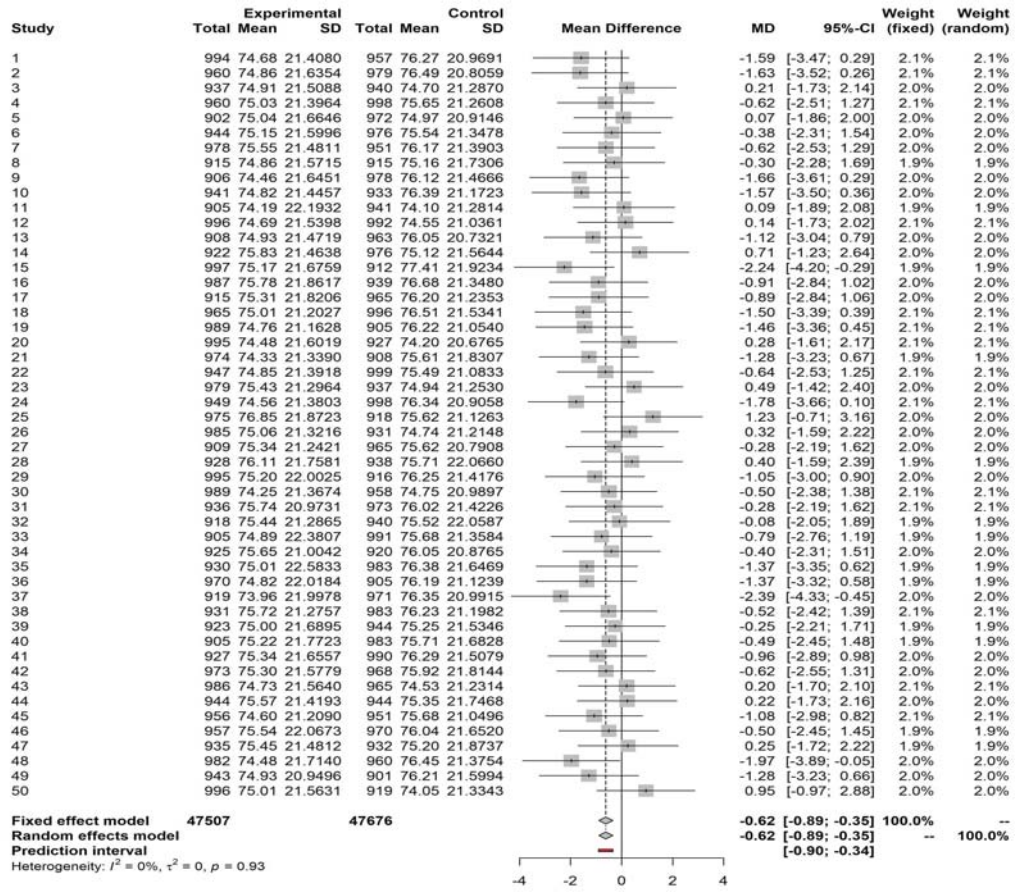

Figura A.11: Meta-análisis con muestras entre 901 y 1,000.

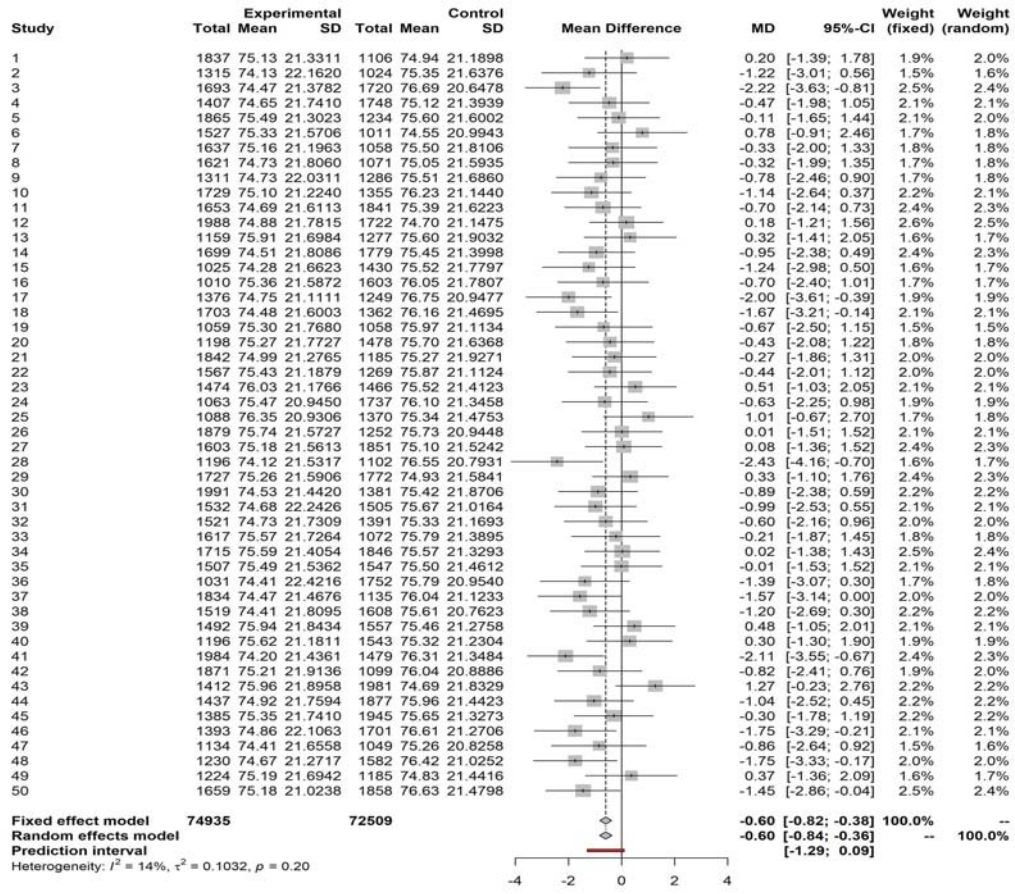

Figura A.12: Meta-análisis con muestras entre 1,001 y 2,000.

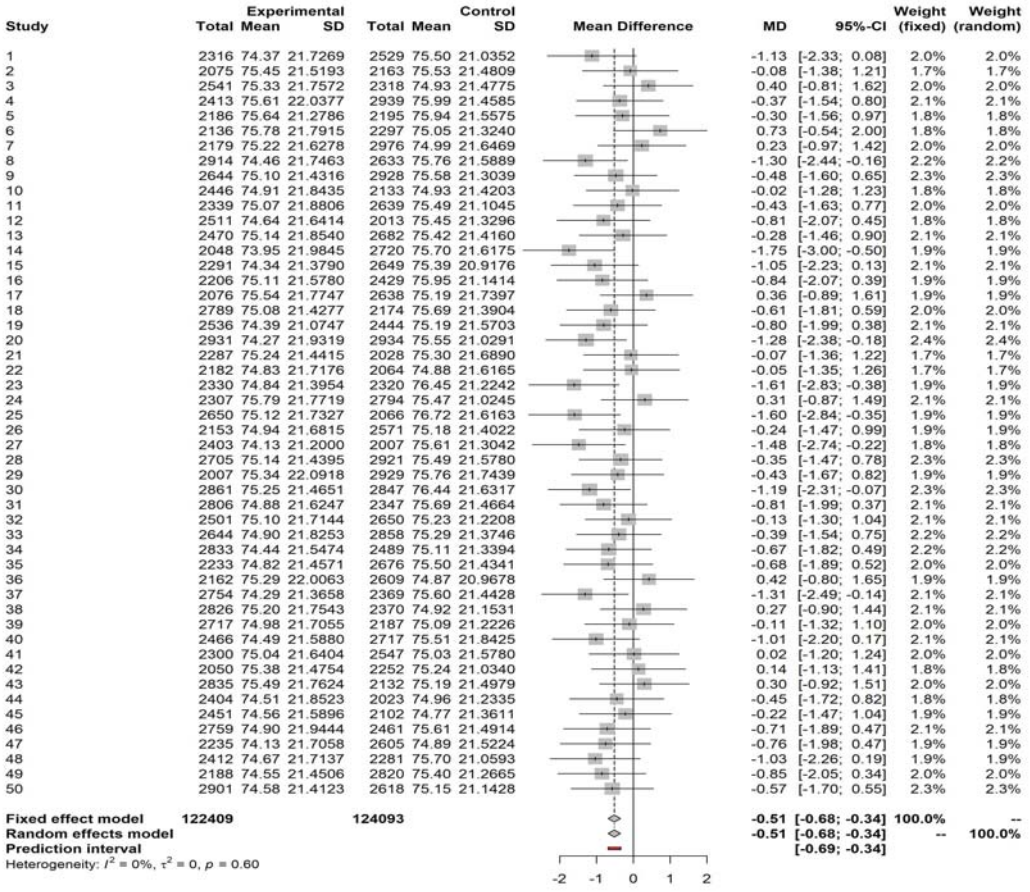

Figura A.13: Meta-análisis con muestras entre 2,001 y 3,000.

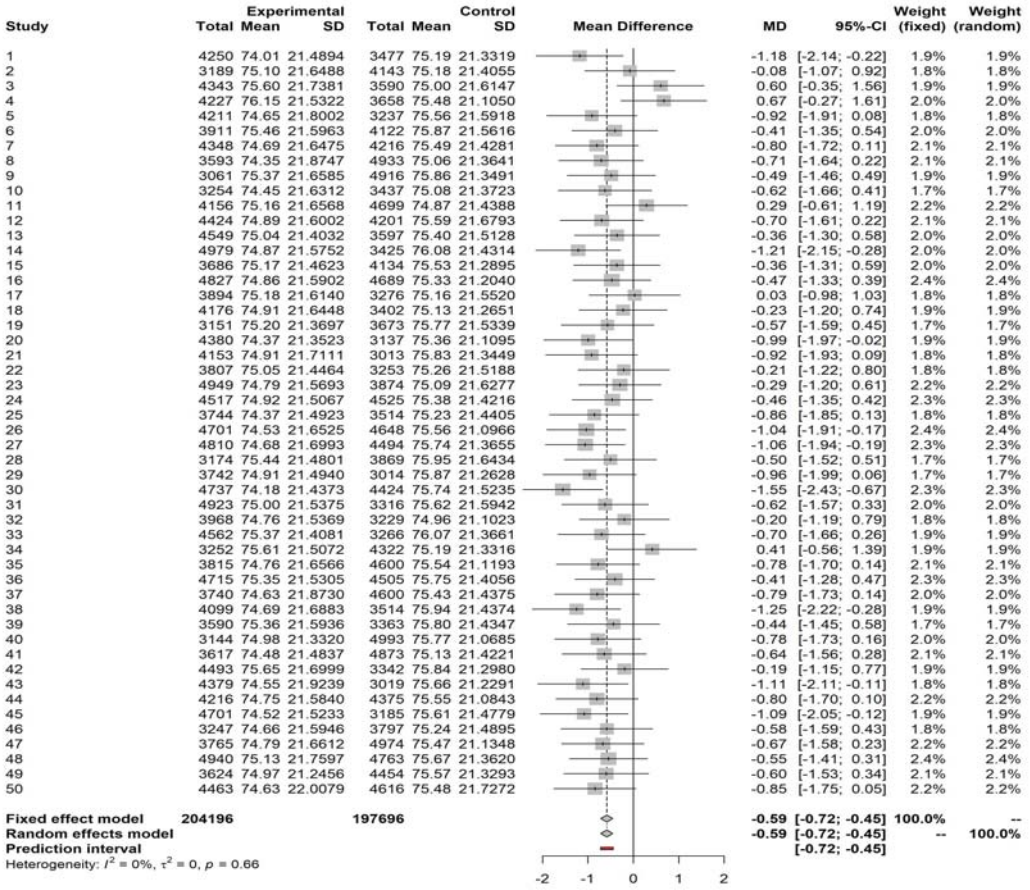

Figura A.14: Meta-análisis con muestras entre 3,001 y 5,000.

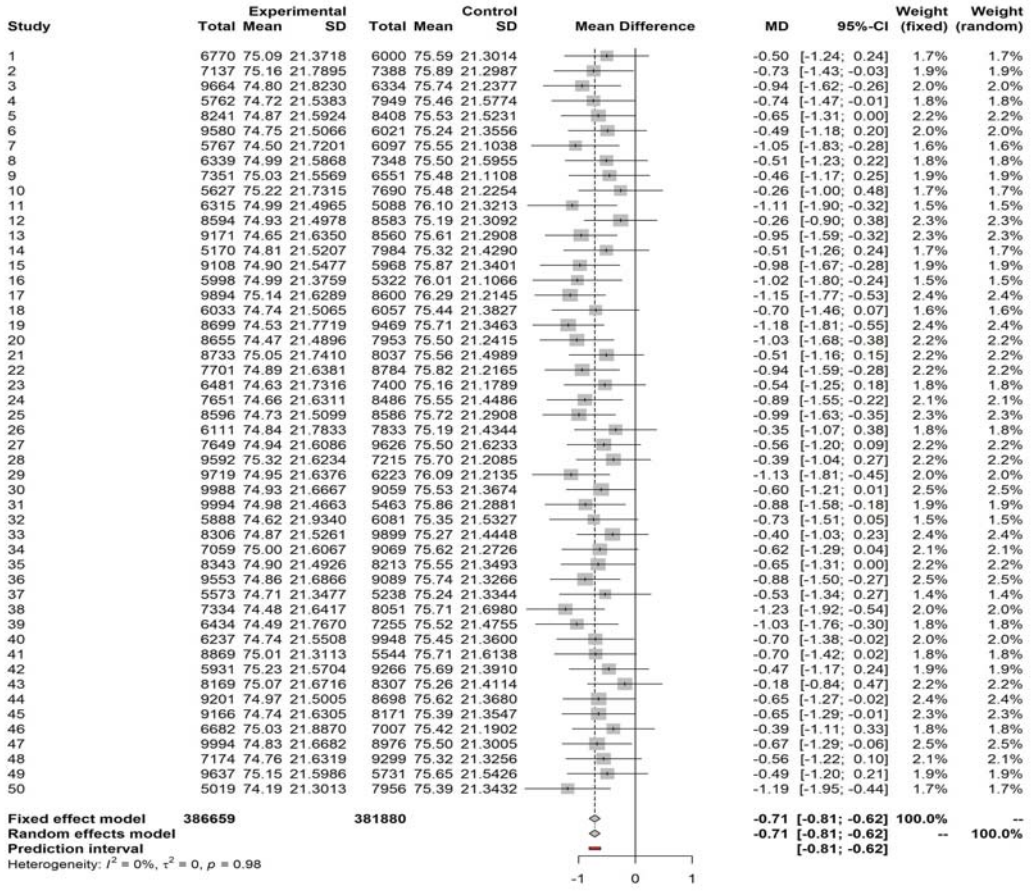

Figura A.15: Meta-análisis con muestras entre 5,001 y 10,000.

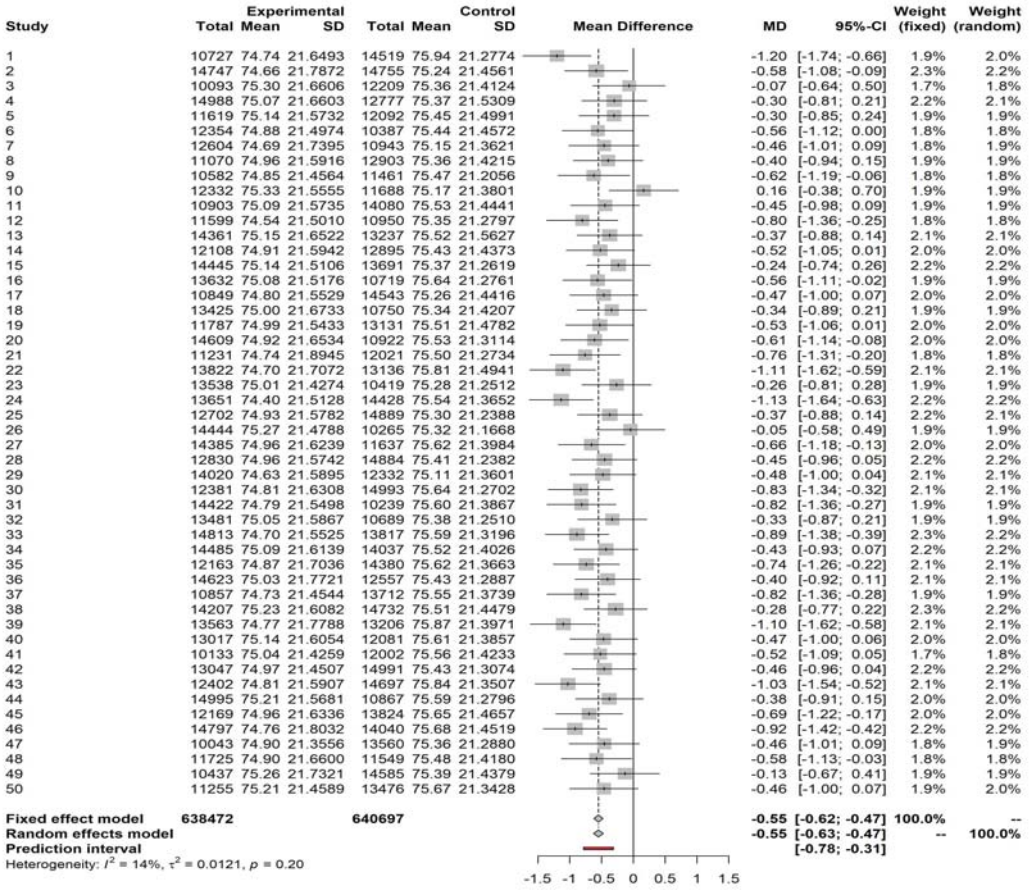

Figura A.16: Meta-análisis con muestras entre 10,001 y 15,000.

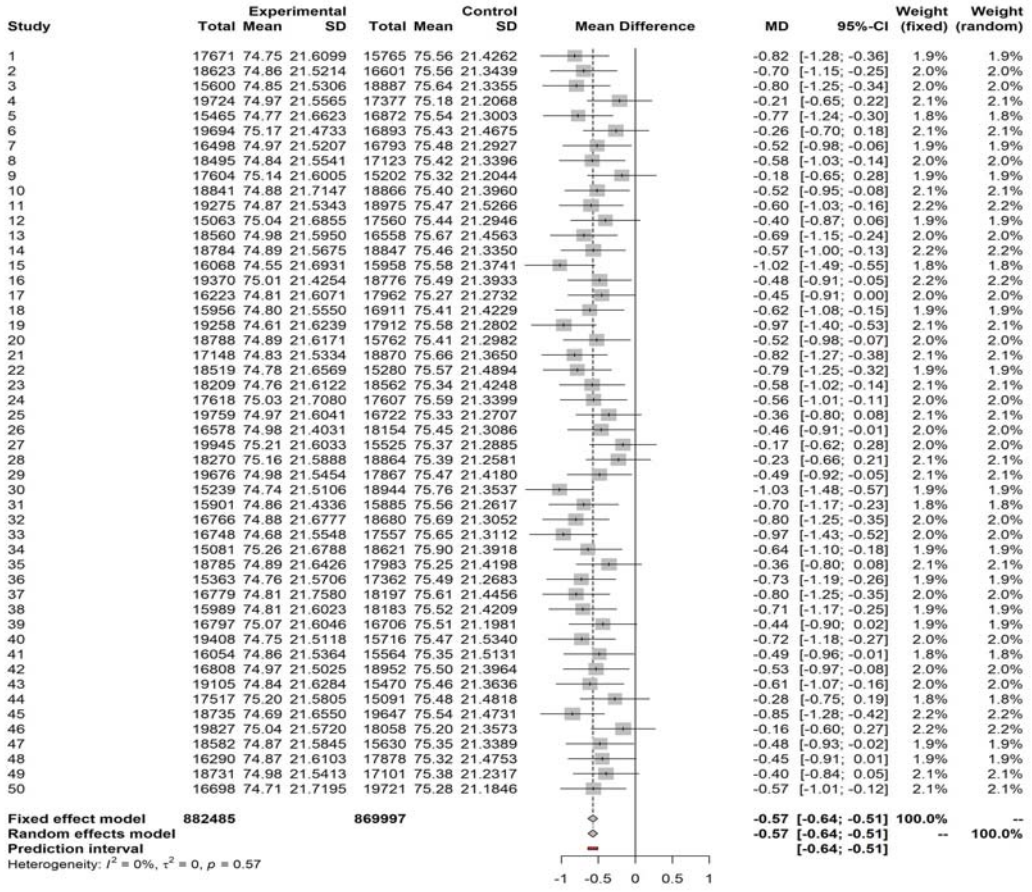

Figura A.17: Meta-análisis con muestras entre 15,001 y 20,000.

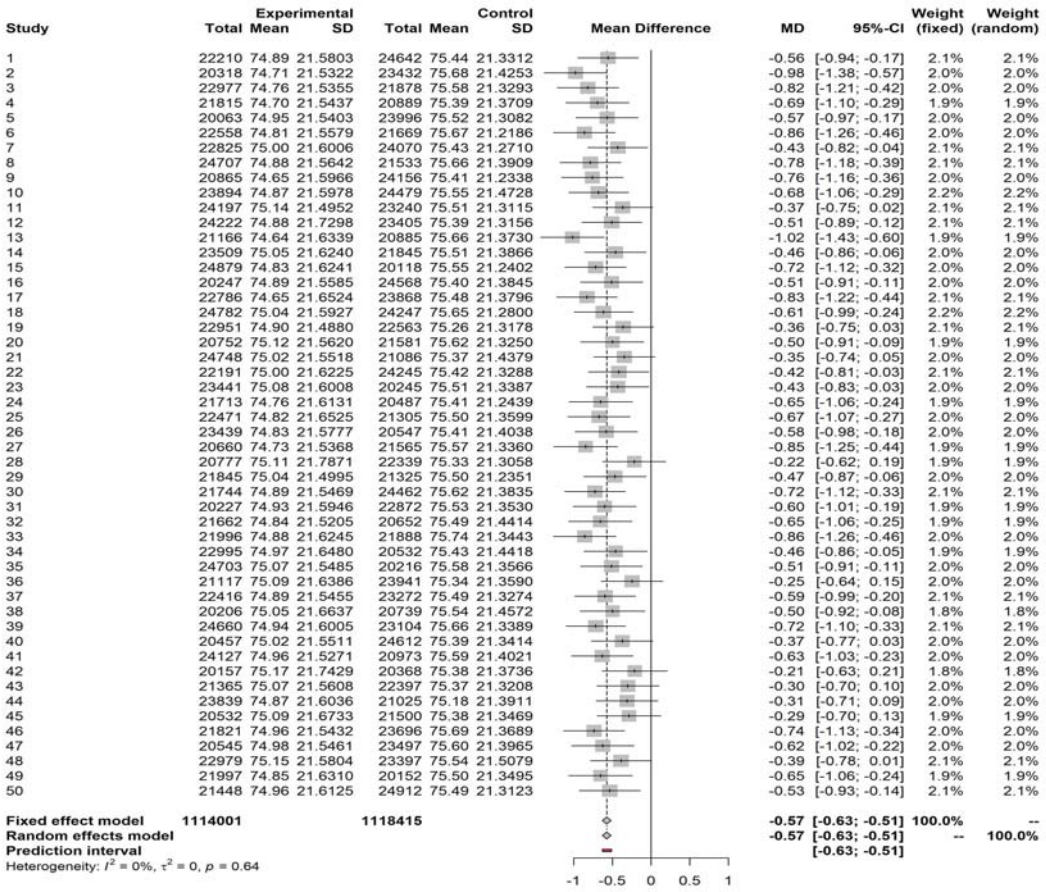

Figura A.18: Meta-análisis con muestras entre 20,001 y 25,000.

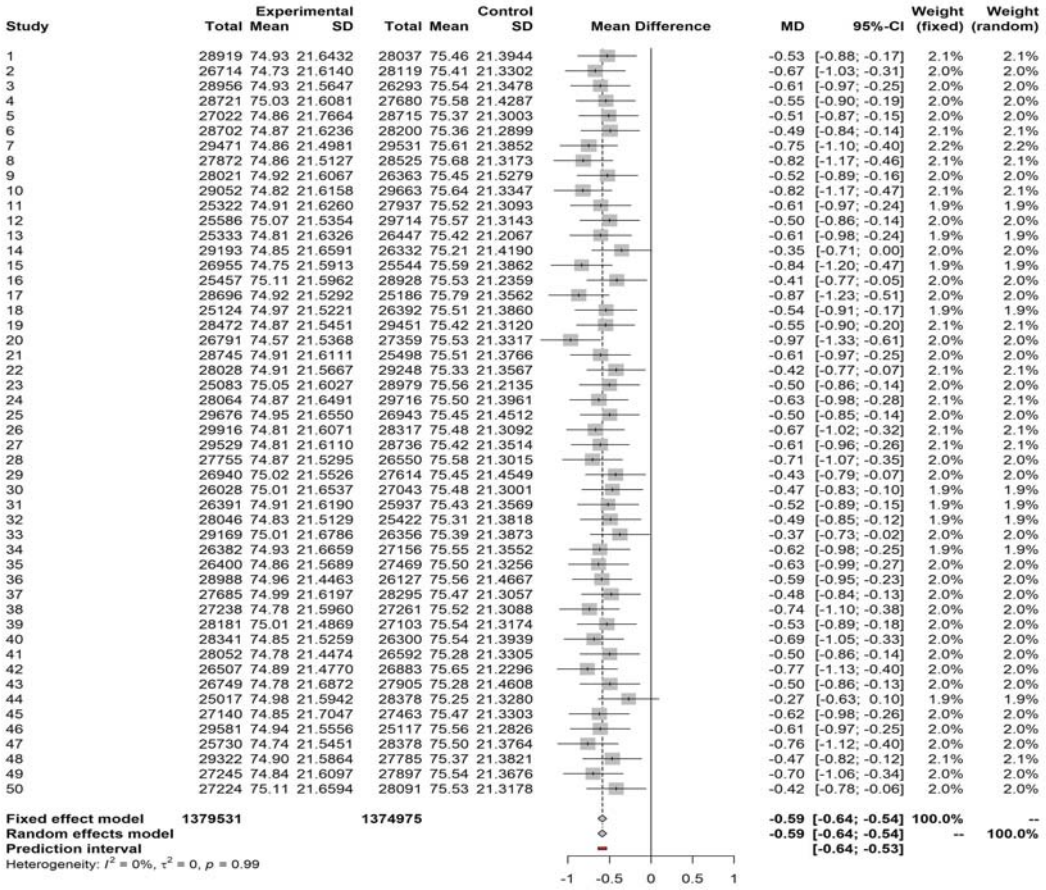

Figura A.19: Meta-análisis con muestras entre 25,001 y 30,000.

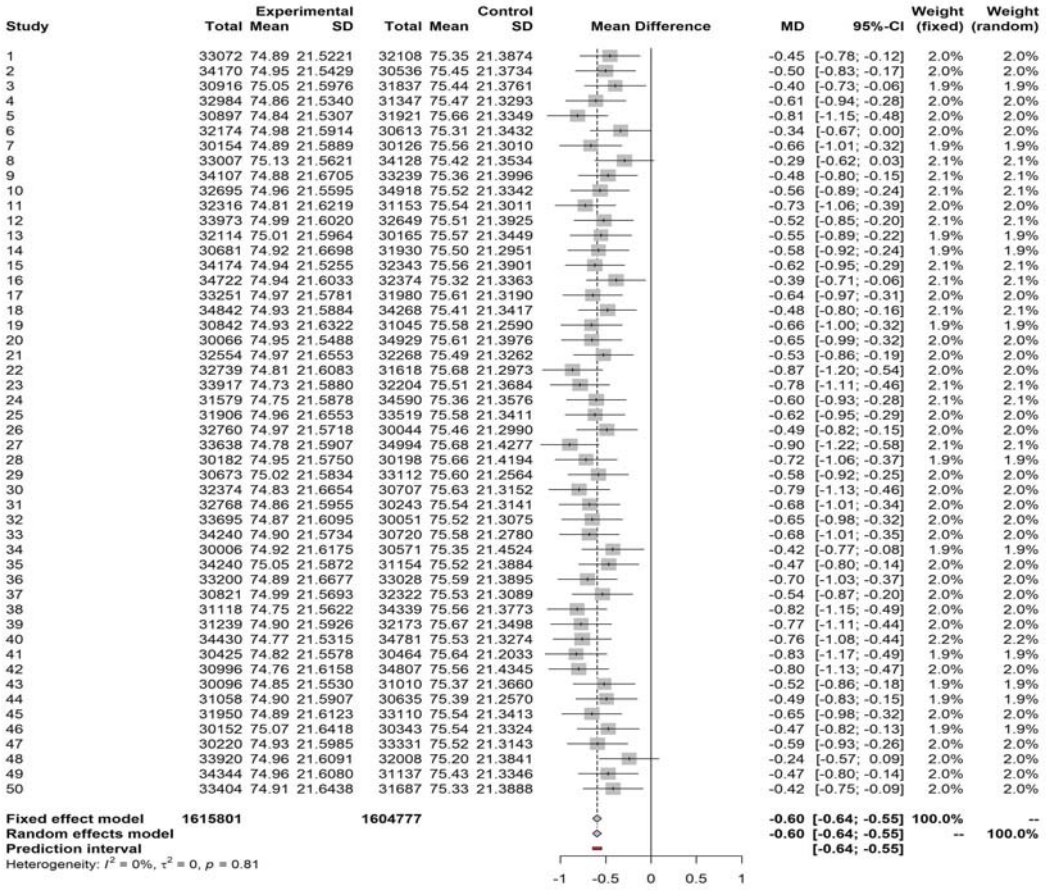

Figura A.20: Meta-análisis con muestras entre 30,001 y 35,000.

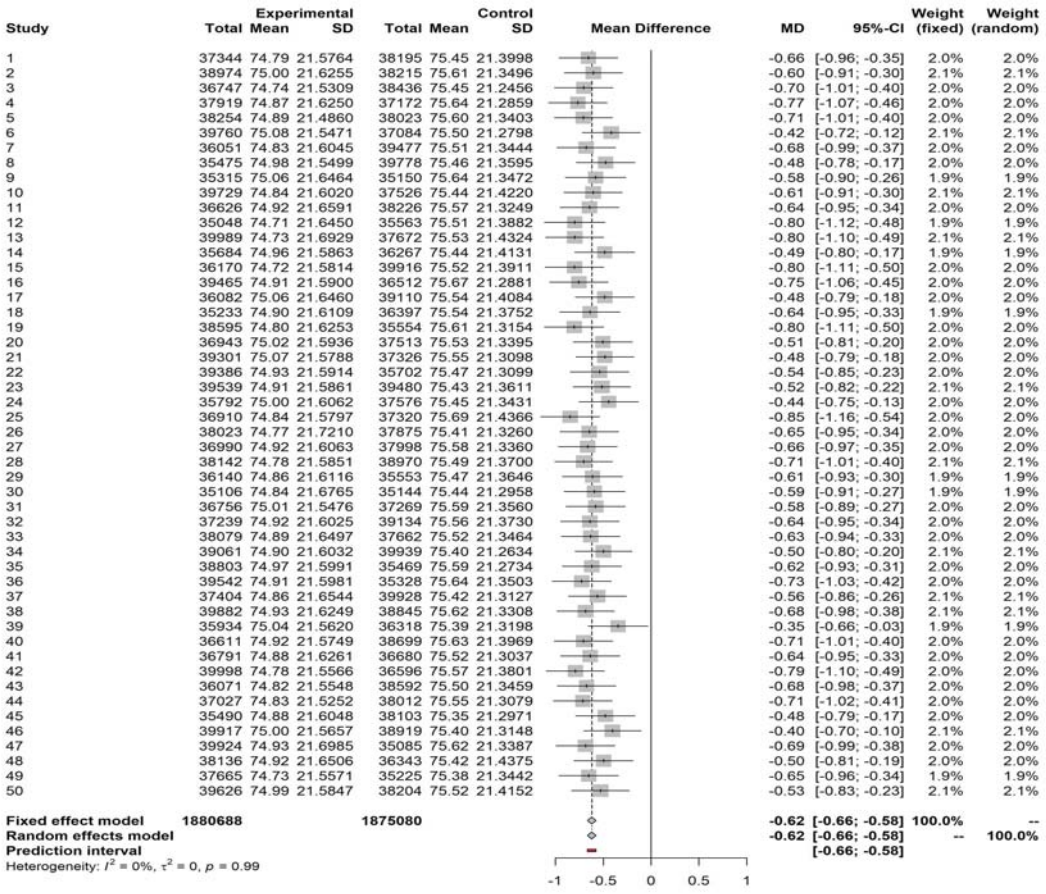

Figura A.21: Meta-análisis con muestras entre 35,001 y 40,000.

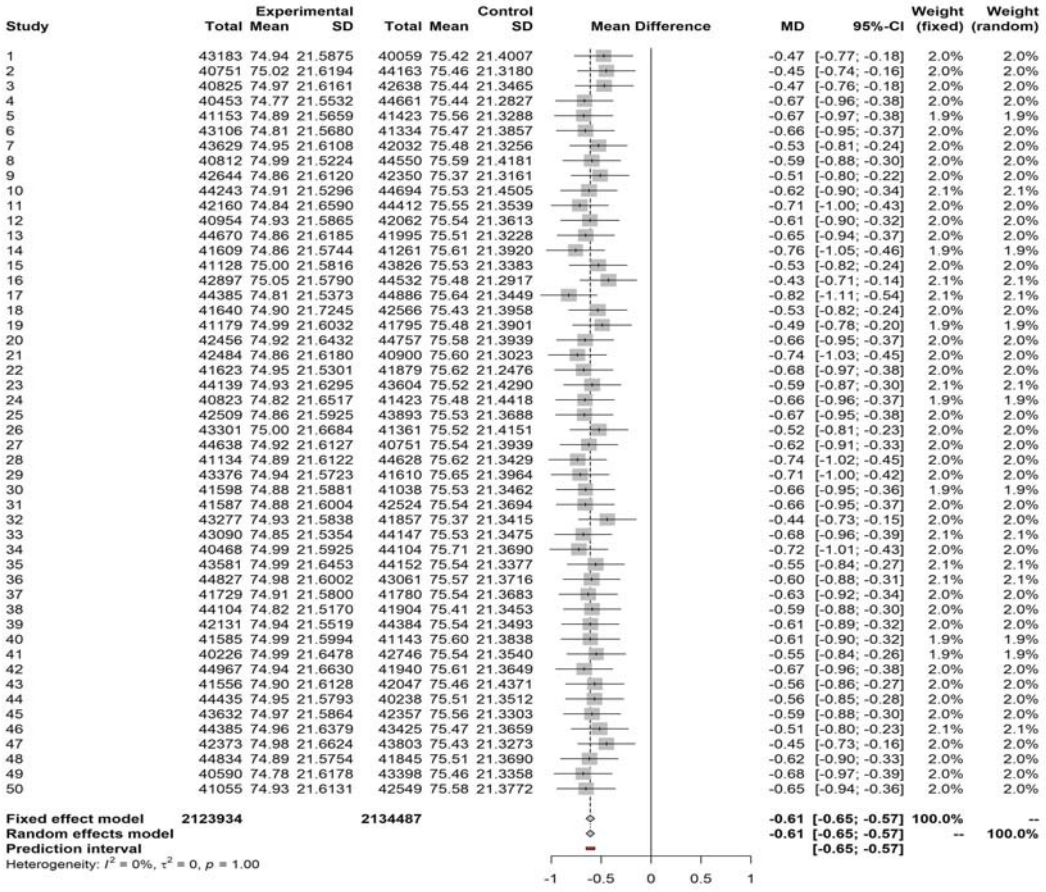

Figura A.22: Meta-análisis con muestras entre 40,001 y 45,000.

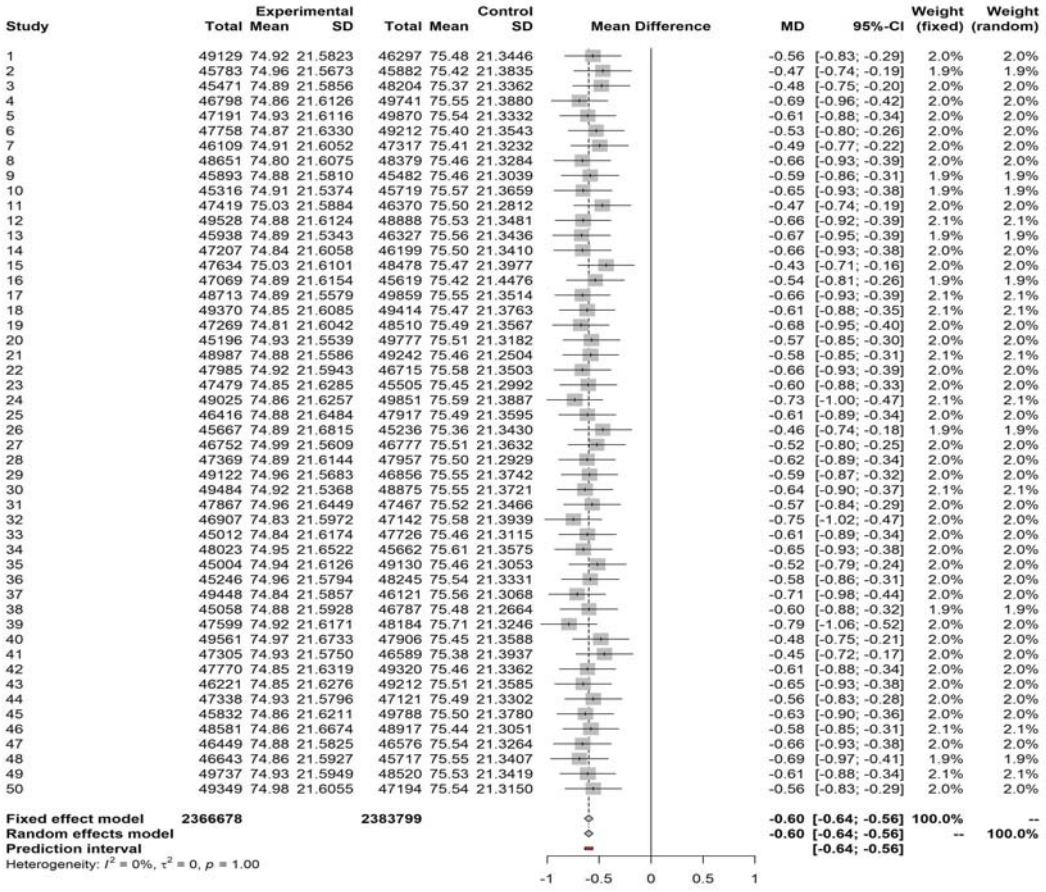

Figura A.23: Meta-análisis con muestras entre 45,001 y 50,000.

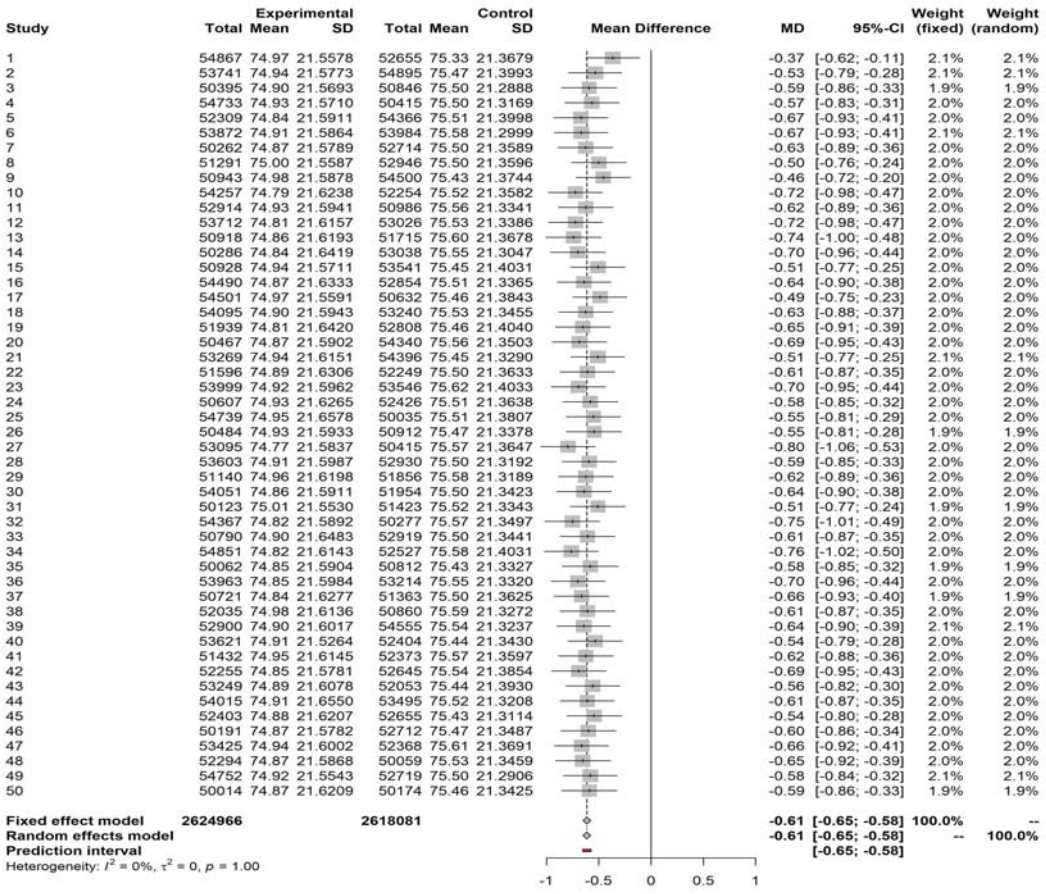

Figura A.24: Meta-análisis con muestras entre 50,001 y 55,000.

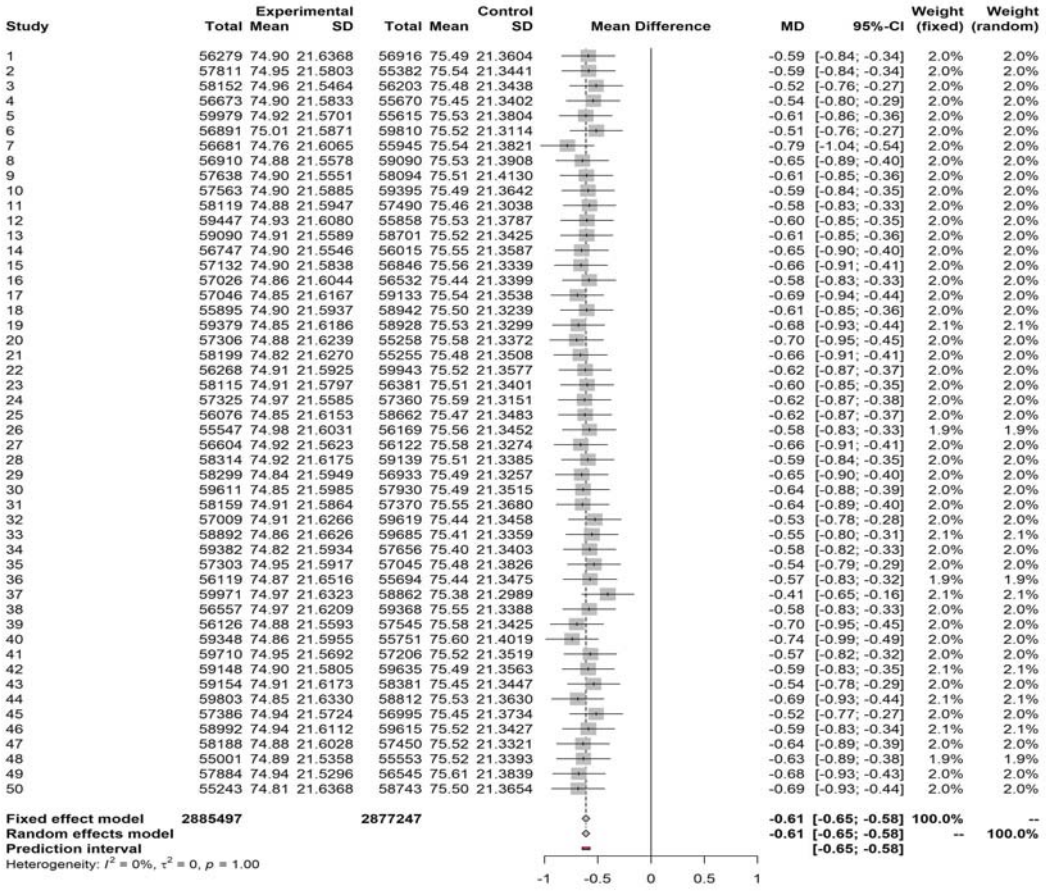

Figura A.25: Meta-análisis con muestras entre 55,001 y 60,000.

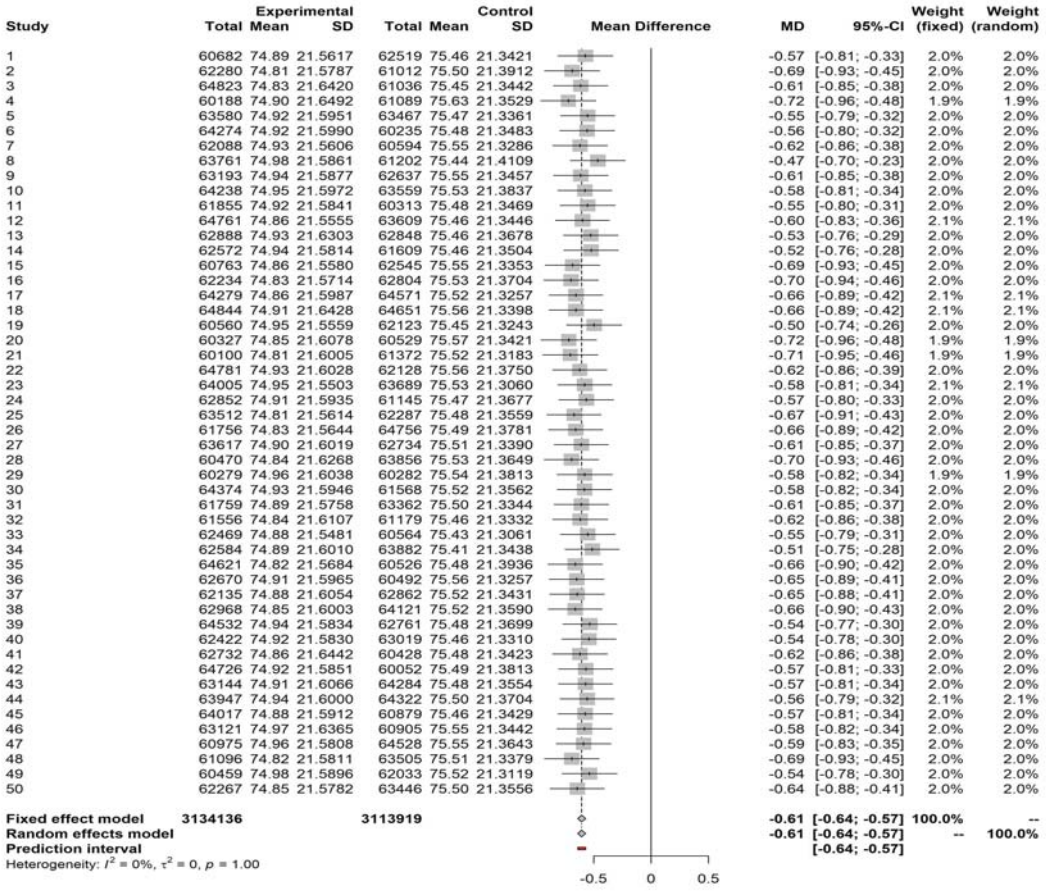

Figura A.26: Meta-análisis con muestras entre 60,001 y 65,000.

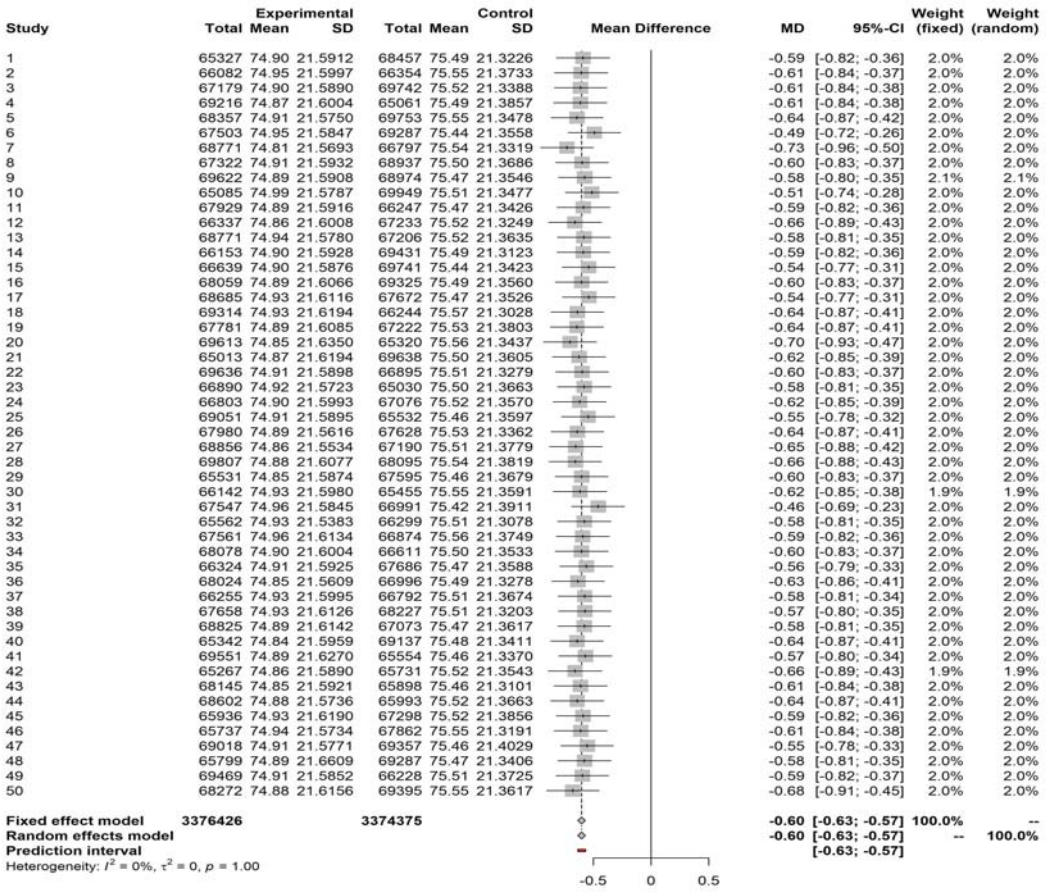

Figura A.27: Meta-análisis con muestras entre 65,001 y 70,000.

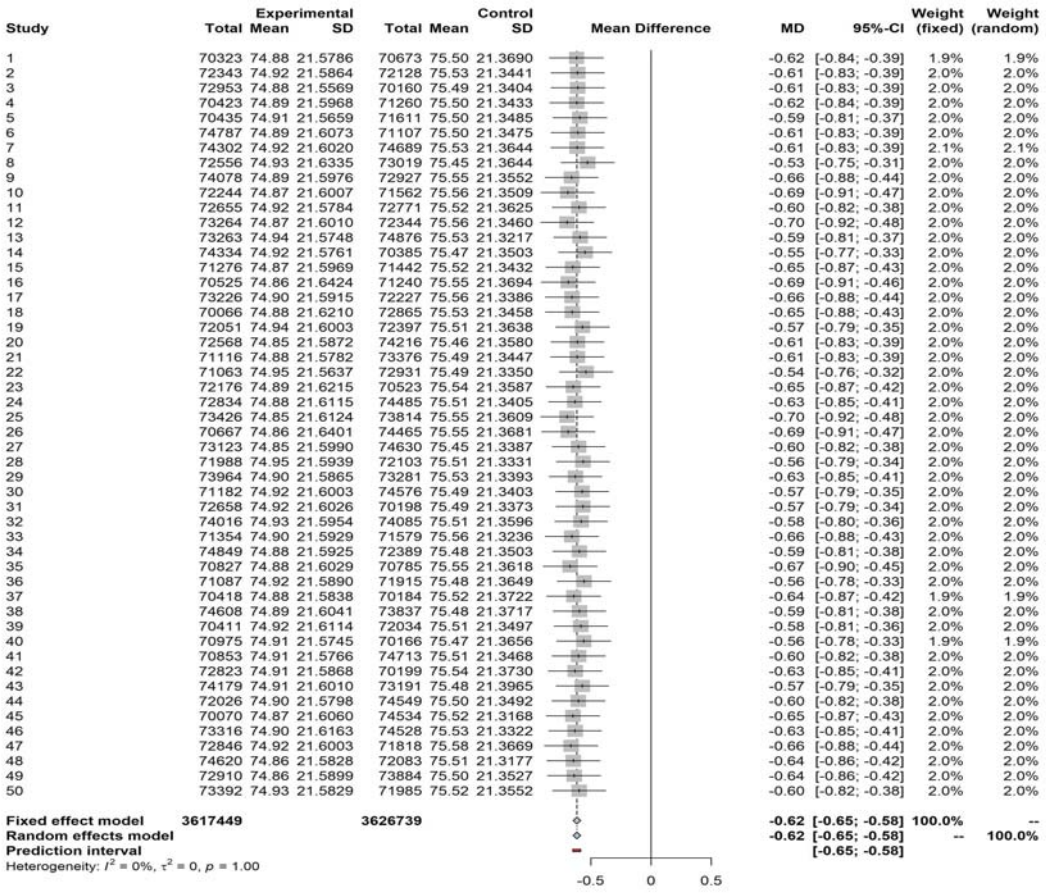

Figura A.28: Meta-análisis con muestras entre 70,001 y 75,000.

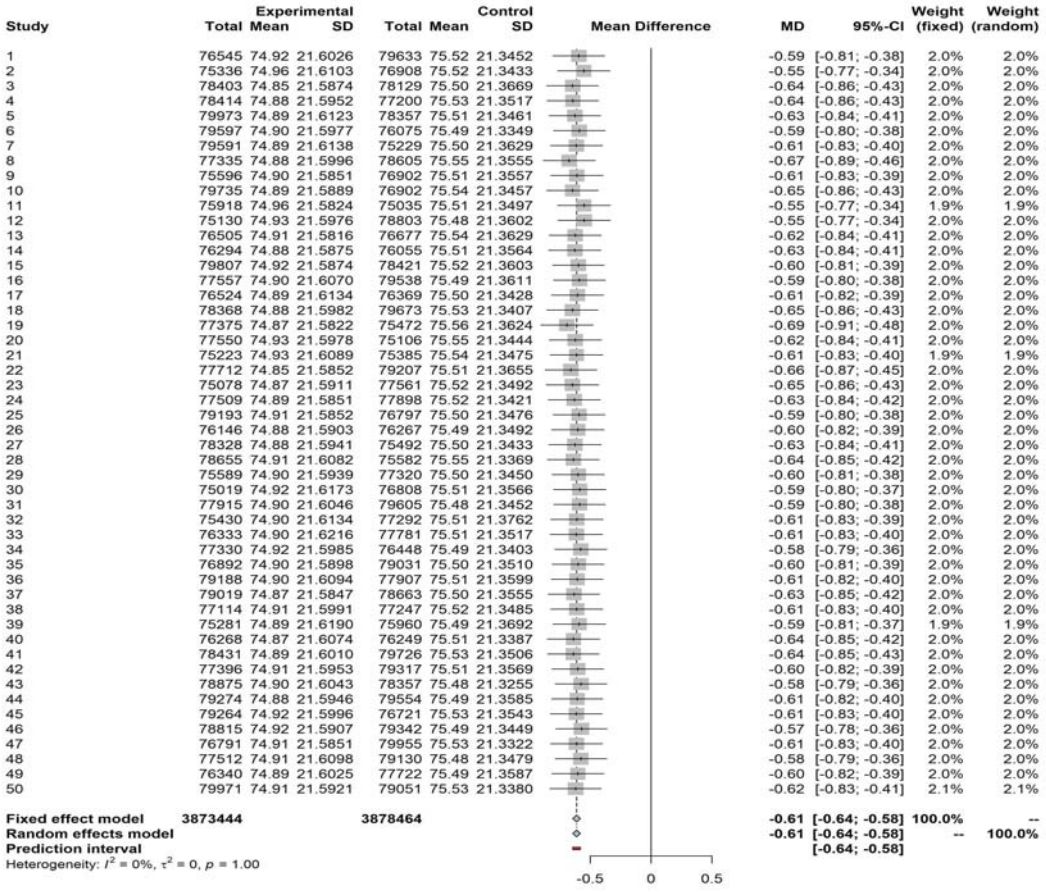

Figura A.29: Meta-análisis con muestras entre 75,001 y 80,000.

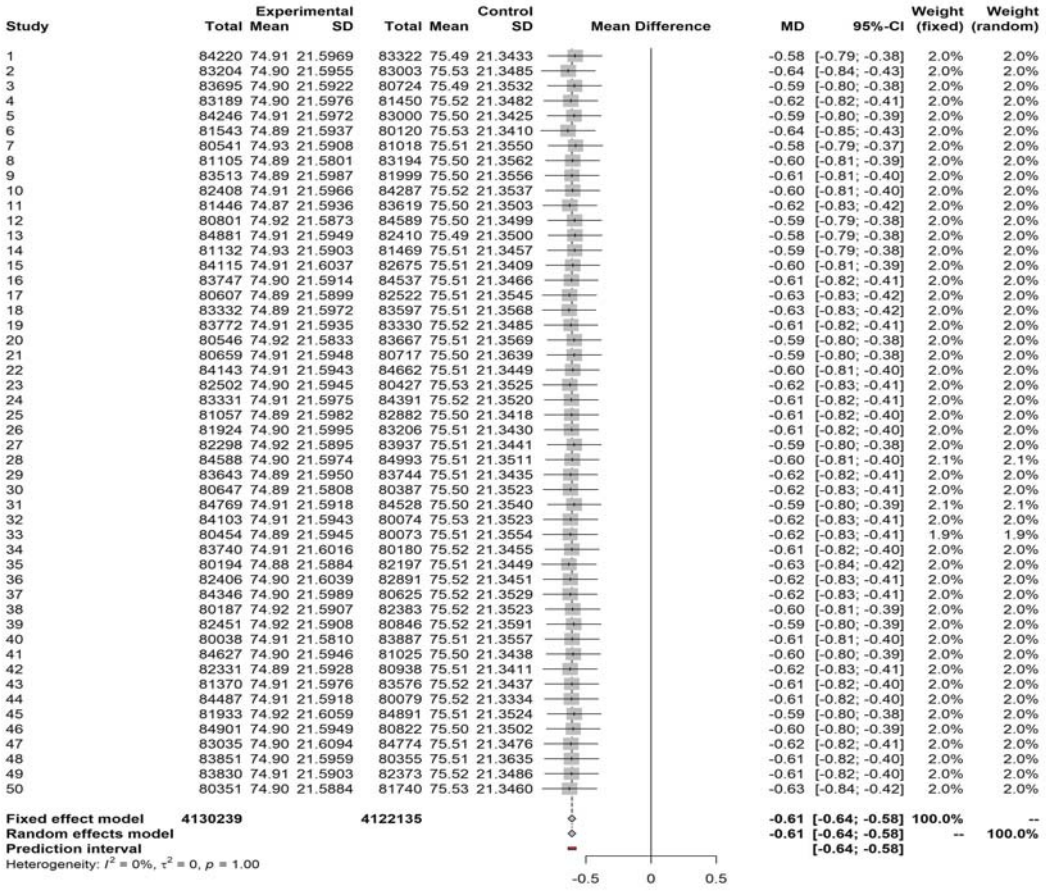

Figura A.30: Meta-análisis con muestras entre 80,001 y 85,000.

Apéndice B

Anexo 2. Códigos en R

B.1. Código para generar el arreglo para realizar los meta-análisis

```
 require(doBy)
require(meta)
library(meta)

# Se definen las rutas de entrada y de salida

ent1 <- "C:/ENT/EX1.csv"
ent2 <- "C:/ENT/EX2.csv"
sal <- "C:/SAL/Meta2"

# Se leen las poblaciones

pob2<-read.csv(ent1,as.is=TRUE)$EX1
pob1<-read.csv(ent2,as.is=TRUE)$EX2

#A continuación se definen grupos de muestras

gpoinf<-c(2,11,21,31,51,101,501,601,701,801,901,1001,2001,
3001,5001,10001,15001,20001,25001,30001,35001,40001,45001,
```

```
50001,55001,60001,65001,70001,75001,80001) #Límite inferior

gposup<-c(10,20,30,50,100,500,600,700,800,900,1000,2000,
3000,5000,10000,15000,20000,25000,30000,35000,40000,45000,
50000,55000,60000,65000,70000,75000,80000,85000)  #Límite superior
grupos<-length(gpoinf)


#Se definen las repeticiones

repeticiones<-c(50)

#Información a generar

informacion<-c("K","Grupo","n1","M1","S1","n2","M2","S2","S","d","t",
"p-value","Rechazo")
infometa<-c("Repeticion","Grupo","n1","M1","S1",
"n2","M2","S2")#Información para el archivo de meta-análisis

#Se definen las funciones a emplear en los resumenes

sumfun <- function(x, ...)\{c(m=mean(x, na.rm=TRUE, ...),
s=sd(x, na.rm=TRUE, ...), mi=min(x,na.rm = TRUE),mx=max(x,na.rm = TRUE),
sm<-sum(x,na.rm = TRUE))}

#Se crea un arreglo donde se depositarán los resúmenes para el meta-análisis

nam3<-"meta.csv"
salida3<-paste(sal,nam3,sep="/")

meta<-matrix(NA,nrow = grupos*length(repeticiones),ncol = 8)
colnames(meta)<-infometa
```

```
#Se procede a realizar las simulaciones

for(i in 1:length(repeticiones)){
  nam<-paste(paste("Simulacion",repeticiones[i],sep = ""),"csv",sep = ".")
  nam2<-"concentrado.csv"
  salida<-paste(sal,nam,sep="/")
  salida2<-paste(sal,nam2,sep="/")
  sim<-matrix(0,nrow=(repeticiones[i]*grupos),ncol = 13)
  colnames(sim)<-informacion
  for(j in 1:grupos){
  nam<-paste(paste("Grupo",j,sep = ""),"csv",sep = ".")
  n1<-rep(NA,repeticiones[i])
  #Vector con los tamaños de muestra de la primera población
  n2<-rep(NA,repeticiones[i])
  #Vector con los tama\~nos de muestra de la segunda población
  media1<-rep(NA,repeticiones[i])
  #Vector con las medias de la primera población
  media2<-rep(NA,repeticiones[i])
  #Vector con las medias de la segunda población
  desv1<-rep(NA,repeticiones[i])
  #Vector con las desviaciones de la primera población
  desv2<-rep(NA,repeticiones[i])
  #Vector con las desviaciones de la segunda población
    for(k in 1:repeticiones[i]){
    tm1<-round(runif(1,gpoinf[j],gposup[j]),0) #tamaño de muestra1
    tm2<-round(runif(1,gpoinf[j],gposup[j]),0) #tamaño de muestra2
    n1[k]<-tm1
    n2[k]<-tm2
    muestra1<-sample(pob1,tm1) #Muestra de la primera población
    muestra2<-sample(pob2,tm2) #Muestra de la segunda población
    muestra<-c(muestra1,muestra2)
    prueba<-t.test(muestra1,y=muestra2)
    sim[(k+(j-1)*repeticiones[i]),1]<-k
    sim[(k+(j-1)*repeticiones[i]),2]<-j
```

```
 sim[(k+(j-1)*repeticiones[i]),3]<-tm1
 sim[(k+(j-1)*repeticiones[i]),4]<-mean(muestra1,na.rm = TRUE)
 sim[(k+(j-1)*repeticiones[i]),5]<-sd(muestra1,na.rm = TRUE)
 sim[(k+(j-1)*repeticiones[i]),6]<-tm2
 sim[(k+(j-1)*repeticiones[i]),7]<-mean(muestra2,na.rm = TRUE)
 sim[(k+(j-1)*repeticiones[i]),8]<-sd(muestra2,na.rm = TRUE)
 sim[(k+(j-1)*repeticiones[i]),9]<-sd(muestra,na.rm = TRUE)
 sim[(k+(j-1)*repeticiones[i]),10]
<-abs(mean(muestra1,na.rm = TRUE)-mean(muestra2,na.rm = TRUE))/sd(muestra,na.rm = TRU
  sim[(k+(j-1)*repeticiones[i]),11]<-prueba #statistic
  sim[(k+(j-1)*repeticiones[i]),12]<-prueba #p.value
  if(prueba\$p.value<0.05)sim[(k+(j-1)*repeticiones[i]),13]<-1
   media1[k]<-mean(muestra1,na.rm = TRUE)
   media2[k]<-mean(muestra2,na.rm = TRUE)
   desv1[k]<-sd(muestra1,na.rm = TRUE)
   desv2[k]<-sd(muestra2,na.rm = TRUE)
}
 pond1<-n1/sum(n1) #Vector de ponderadores1
 pond2<-n2/sum(n2)  #Vector de ponderadores2
 n1pond<-sum(pond1*n1) #n1 ponderado
 n2pond<-sum(pond2*n2) #n2 ponderado
 media1pond<-sum(pond1*media1) #media1 ponderada
 media2pond<-sum(pond2*media2) #media2 ponderada
 desv1pond<-sum(pond1*desv1) #desviación1 ponderada
 desv2pond<-sum(pond1*desv2) #desviación2 ponderada
 fila<-(i-1)*grupos+j
 meta[fila,1]<-i
 meta[fila,2]<-j
 meta[fila,3]<-n1pond
 meta[fila,4]<-media1pond
 meta[fila,5]<-desv1pond
 meta[fila,6]<-n2pond
meta[fila,7]<-media2pond
meta[fila,8]<-desv2pond
```

```
 }
 tbl<-summaryBy(cbind(M1,M2,d,t,p.value,Rechazo)~Grupo,data=data.frame(sim),FUN = sumfun
 write.csv(sim,salida,row.names = FALSE,na=" ")
 write.table(paste("K=",repeticiones[i],sep=""),salida2,row.names = FALSE,na=" ",append
 write.table(rep("",2),salida2,row.names = FALSE,na=" ",append = TRUE, sep=",")
 write.table(tbl,salida2,row.names = FALSE,na=" ",append = TRUE, sep=",")
}
write.csv(meta,salida3,row.names = FALSE,na=" ")
```

B.2. Código para realizar los gráficos de meta-análisis

```
library(meta)

arch<-"C:/SAL/Meta2/Simulacion50.csv"
sal<-"C:/SAL/Meta2"
X<-read.csv(arch,as.is = TRUE)

for(i in 1:30){
Y<-subset(X,Grupo==i,select=c(K,n1,M1,S1,n2,M2,S2))
m2 <- metacont(n1, M1, S1,
               n2, M2, S2,
               data = Y, studlab = K,prediction=TRUE)

fig<-paste(paste("Figura",(2+i),sep=""),"jpg",sep = ".")
imagen<-paste(sal,fig,sep="/")
jpeg(imagen,width = 40,height = 50,units = "cm",res = 300,pointsize = 12)
forest(m2)
dev.off()
}
```

Referencias

[1] Callegaro, A. Ndour, C. Aris, E. Legrand, C. (2018). A note on tests for relevant differences with extremely large sample sizes. *Biometrical Journal*. DOI: 10.1002/bimj.201800195.

[2] Campillo-Labrandero, M. Martínez-González, A. García-Minjares, M. Guerrero-Mora, L. Sánchez-Mendiola, M.(2019). Desempeño académico y egreso en 25 generaciones de estudiantes de la Facultad de Medicina de la UNAM. *Revista Educación Médica*; doi:10.1016/j.edumed.2019.05.003.

[3] Casella, G. Berger, R. L.(1990). *Statistical Inference*. Duxbury Press. Belmont, California. 220-228.

[4] Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*. 65, 145–153.

[5] Cohen, J. (1990). Things I Have Learned (So Far). *American Psychologist*. Vol. 45. No 12. 1304-1312.

[6] Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*. Vol. 1, No. 3, 98-101.

[7] Coronel-Nuñez, Z. Ruiz Velasco Acosta, S. (2019). *El tamaño del efecto*. Tesis. UNAM.

[8] Dirección General de Administración Escolar (DGAE) UNAM. (2020). *Demanda de ingreso al bachillerato 2000-2020*.

[9] Dirección General de Planeación UNAM. (2020). *Agenda Estadística de la UNAM 2020*. México.

[10] Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*.1992;268(17):2420-5.

[11] Hodges, J. L., Lehmann, E. L. (1954). Testing the approximate validity of statistical hypothesis. *Journal of the Royal Statistical Society, Series B.*,16, 262–268.

[12] Gross, J. Ligges, U. (2015). *Package nortest*. R project. Recuperado de https://cran.r-project.org/web/packages/nortest/nortest.pdf.

[13] Lazzeroni, LC. Lu, Y. Belitskaya-Lévy, I.(2014). P-values in genomics: Apparent precision masks high uncertainty. *Molecular Psychiatry*. 19, 1336–1340; doi:10.1038/mp.2013.184.

[14] Ledesma, R. Macbeth, G. Cortada de Kohan, N. (2008). Tamaño del efecto: revisión teórica y aplicaciones con el sistema estadístico ViSta. *Revista Latinoamericana de Psicología*. Volumen 40, No 3. 425-439.

[15] Lin, M. Lucas, H. Shmueli, G. (2013). Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*. Vol. 24, No. 4, pp. 906-917.

[16] Martínez-González, A. Manzano-Patiño,A. P. García-Minjares, M. Herrera-Penilla, C.J. Buzo-Casanova, E.R. Sánchez-Mendiola, M.(2018). Grado de conocimientos de los estudiantes al ingreso a la licenciatura y su asociación con el desempeño escolar y la eficiencia terminal. Modelo multivariado. *Revista de la Educación Superior*. 47 (188), pp.57-85. ANUIES.

[17] Mayo, D. Cox D.R.(2006). *Frequentist Statistics as a Theory of Inductive Inference*. Lecture Notes-Monograph Series, Vol. 49, Optimality: The Second Erich L. LehmannSymposium, pp. 77-97. Institute of Mathematical Statistics.

[18] Marden, J. (2000). Hypothesis Testing: From p Values to Bayes Factors. *Journal of the American Statistical Association*. Vol. 95, No. 452, pp.1316-1320.

[19] Moncada-Hernández, S.(2014). Cómo realizar una búsqueda de información eficiente. Foco en estudiantes, profesores e investigadores en el área educativa. *Revista Investigación en Educación Médica*. 3(10), pp. 106-115.

[20] Mood, A. Graybill, F. Boes, D. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill. Third edition. Singapore. 401-463

[21] Navarro, P. Ottone, NE. Acevedo, C. Cantín, N. (2017). Pruebas estadísticas utilizadas en revistas odontológicas de la red SciELO. *Avances en Odontoestomatología*. 33 (1): 25-32.

[22] Riley, R.D. Higgins, J. P. T. Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ: British Medical Journal*. Vol. 342, No. 7804, pp. 964-967.

[23] Salgado, J. F. (2018). Transforming the Area under the Normal Curve (AUC) into Cohen's d, Pearson's rpb, Odds-Ratio, and Natural Log Odds-Ratio: Two Conversion Tables. *The European Journal of Psychology Applied to Legal Context*. 10(1) 35-47.

[24] Schwarzer, G. Carpenter,J.R. Rücker, G.(2015). *Meta-Analysis with R*. Springer. First edition. Switzerland pp 21-52.

[25] Thompson, B. (2006). Critique of p-Values. *International Statistical Review / Revue Internationale de Statistique*. Vol. 74, No. 1, pp. 1-14.

[26] Ventura-León, J. (2018). Otras formas de entender la d de Cohen. *Revista Evaluar*. 18(3), 73-78.

[27] Verdam, M. Oort, F. Sprangers, M. (2014). Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research*. Vol. 23, No. 1, pp. 5-7.

[28] Victor, N. (1987). On clinically relevant differences and shifted null hypotheses. *Methods of Information in Medicine*. 26, 155–162.

[29] Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

[30] Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biometrical Journal*. 59(5), 854–872.