Conceptos Básicos de Estadísticas Inferenciales Aplicadas a la Investigación Educativa: Volumen I

Tabla de Contenidos

Lista de Tablas y Figuras

Agradecimientos y Reconocimientos

Mi agradecimiento más profundo va dirigido a mi familia: Laura, Isabella Concepción, Boogie, Simón y Prathiba. También, le doy gracias a mi familia extendida (abuelos, compadres, madres, hermana, primas, primos, tías y tíos) por haberme dado las herramientas para la creación de este libro. Otro agradecimiento va a mis amigos con los que he hablado de intereses académicos además de otras muchas cosas más. Reconozco a las y los Administrativos, Profesores y Alumnos del Instituto de Ciencias Sociales y Administración de la Universidad Autónoma de Ciudad Juárez por su apoyo, conocimientos y cuestionamientos para llevar a cabo este libro. La interacción con ellos y ellas me ha enseñado elementos de estadísticas que habían pasado desapercibidos.

Prologo

Resumen y Descripción de la Obras

En la investigación educativa, se usan las estadísticas y análisis psicométricos para describir datos, entender relaciones entre variables y desarrollar teorías. En este presente libro, solo se cubren los principios de estadísticas inferenciales y sus efectos para dar respuesta a la siguiente pregunta: ¿Qué son las estadísticas inferenciales y cuáles son sus implicaciones? Al darle respuesta, se espera que el panorama se habrá para el lector al poder aplicar una serie de conceptos que se desarrollan en este libro para los análisis estadísticos más sofisticados como: test t de estudiante, análisis de la varianza, correlaciones, chi cuadrada, regresión, análisis multivariado, ecuaciones estructurales, etc., y, también, efectos (e.g., d de Cohen, r de Pearson, f^2 de Cohen, Coeficiente de determinación r^2, V de Cramer, R^2, etc.). En concreto, el objetivo del presente libro es mostrar cómo llevar a cabo test de significancia estadística con sus implicaciones, explicar principios básicos de muestreo, estimar intervalos de confianza para el promedio, y calcular el tamaño de un efecto práctico: d de Cohen.

Los principios que se cubren en este presente libro son aplicables a métodos de investigación cuantitativa como lo son los meta-análisis (para más información al respecto, ver a Cumming, 2013) y minería de datos (ver a ElAtia, Ipperciel, y Zaiane, 2016). Estos dos métodos estadísticos están ayudando a comprender mejor fenómenos educativos y, también, en otras áreas. Por un lado, los meta-análisis ayudan a calcular el efecto de una variable en otra al usar los resultados de cientos o, tal vez, miles de estudios. Por el otro lado, la minería de datos ha sido definida por algunos investigadores como un proceso de encontrar anomalías, patrones y correlaciones con set de datos grandes para hacer predicciones.

Dirigido a Lectores

Este libro está dirigido a investigadores educativos, profesores, y estudiantes quienes vayan a utilizar análisis de significancia estadística, intervalos de confianza y tamaño del efecto (d de Cohen). Aunque el libro trata los temas de comparación de grupos de una manera bastante general, sus principios de tamaño de la muestra, test de significancia estadística (también llamado *Test de Significancia Estadística de la Hipótesis Nula*: *Null*

Hypotesis Significanse Testing), error, y poder estadístico son esenciales para poner a prueba hipótesis. Asimismo, se cubre la posibilidad de generalizar alguna estadística encontrada en una muestra a su población correspondiente por medio del uso de intervalos de confianza. También, se muestra como cuantificar la magnitud de un efecto (e.g., un tratamiento para mejorar el aprendizaje). En síntesis, este libro tiene como objetivo el ser una herramienta y referencia a la hora de usar test de significancia estadística, intervalos de confianza y efecto.

Conocimientos Previos

Se asume que el usuario o usaría de este libro está familiarizado con estadísticas descriptivas, probabilidad a un nivel elemental, algebra básica, y geometría lineal también a nivel básico. También, ayudaría tener nociones de cómo elaborar preguntas de investigación en investigación educativa cuantifica. Si este no es el caso se recomienda a Ercikan y Roth (2006) y Ponce, Domínguez, Arriaga (2016) para la formulación de preguntas de investigación en la investigación educativa.

Los programas que se usaron para los cálculos del presente libro son Excel y *Gpower*. Este último se puede descargar gratuitamente. Los demás cálculos se llevaron a cabo con una calculadora y con una Tabla de valores *z* que aparece en el Apéndice. Al final de cada capítulo, están los temas de **Preguntas para Reflexionar** y **Opinión del Autor**. La sección de preguntas es para tratar de hacer pensar e investigar al lector o lectora acerca de temas relacionados con el contenido de los capítulos. La opinión de autor es la sección que enfatiza las partes más importantes del capítulo de una manera coloquial.

Otros programas que son muy útiles para las estadísticas son *SPSS*, *Minitab* y *SAS* por el lado de los softwares comerciales. Existen en *You Tube* un gran número de tutoriales para llevar a cabo desde análisis descriptivos hasta inferenciales con estos programas comerciales. Por el lado de acceso gratuito, está el programa llamado *Proyecto R para Cálculos Estadísticos* (*The R Project for Statistical Computing*; https://www.r-project.org/). También, se le conoce simplemente como *R*. Este programa usa código, así que hace su uso un tanto más complicado que los anteriores. Por otro lado, existen cursos en idioma español en el portal *Coursera* (https://es.coursera.org/) para aprender el código que maneja *R*.

Otro recurso para realizar análisis estadísticos es una calculadora para estimar la probabilidad calculada de algún test estadístico. Un ejemplo de lo anterior es la *Calculadora Rápida para Estimar la Probabilidad Calculada* (*Quick P Value Calculator*; https://www.socscistatistics.com/pvalues/Default.aspx). Esta calculadora es muy útil cuando se estima manualmente un test con un coeficiente *z*, *t*, chi cuadrada, *F* o *r*. Simplemente, se introduce estos coeficientes con el nivel del valor del alfa seleccionado del test para calcular el área bajo una curva que representa la probabilidad calculada (i.e., *p* valor o coeficiente de probabilidad calculado). En ciertos casos, es necesario también introducir los grados de libertad y si el test es de una o dos colas. Algunos de estos detalles de los test antes mencionados serán explicados dentro de este presente libro. Por el momento, esta calculadora es un elemento esencial para complementar análisis estadísticos que se elaboren manualmente.

Aporte principal de la Obra

Una de las aportaciones de la obra es el uso de idioma español para los y las personas que así lo prefieren por alguna razón ya que muchas de las referencias están escritas originalmente en el idioma Ingles. Asimismo, se explican las estadísticas inferenciales y el efecto para resolver preguntas de investigación en la educación. Otra de las aportaciones es que se puede usar esta obra para entender análisis más complejos en análisis de multivariado tanto en cuando significancia estadística y sus implicaciones como el efecto. En algunas ocasiones, los libros de investigación educativa son muy extensos porque tratan diferentes tipos de investigación como: la cualitativa, cuantitativa, mixta, investigación histórica, e investigación acción, pero no suelen cubrir tan ampliamente como este presente libro aspectos de poder estadístico, intervalos de confianza y efecto.

Razón para la Creación y Publicación de la Obra

La obra surgió primero como *Material de Apoyo Didáctico* para las clases de métodos de investigación educativa y para seminarios de escritura de tesis. De hecho, el contenido del presente libro se estuvo perfeccionando durante varios semestres en las clases antes mencionadas. Luego, se pensó que esta obra debería de ser de Divulgación ya que en varias universidades y programas se tratan temas de investigación educativa y estadísticas inferenciales.

Objetivos de Aprendizaje del Libro.

Al terminar el libro, el lector será capaz de:

- Comprender los principios de la curva normal.
- Utilizar la Distribución Normal Estandariza como modelo para resolver problemas de probabilidad y estadísticas.
- Entender los análisis estadísticos inferenciales y el Teorema de Tendencia Central para hacer inferencias de las muestras a la población de interés.
- Comprender y calcular diferentes tipos de muestreos e intervalos de confianza.
- Diferenciar entre estadísticas paramétricas y no-paramétricas.
- Formular hipótesis nula y alternativa para poblaciones y muestras.
- Entender los Errores Tipo I y II cuando se trata de generalizar.
- Interpretar los resultados significativos y no significativos de análisis.
- Identificar la diferencia de usar test de una cola con respecto a dos colas y el poder estadístico.
- Entender y calcular el tamaño del efecto llamado *d* de Cohen.

Términos Clave.

- Curva Normal
- Medidas de Tendencia Central
- Curtosis y simetría
- Curva Normal Estandarizada
- Estadísticas Inferenciales
- Teorema de Tendencia Central
- Distribución de Muestras
- Tipos de muestreos
- Intervalos de Confianza
- Estadísticas Paramétricas
- Estadísticas No-paramétricas
- Variable Dependiente
- Variable Independiente
- Hipótesis Nula
- Hipótesis Alternativa
- Error Tipo I y II
- Poder Estadístico
- Test de una Cola y Dos colas
- Tamaño del Efecto
- *d* de Cohen

Capítulo 1. Distribución Normal.

La distribución normal es un modelo que representa algún fenómeno. Como todo modelo tiene sus limitaciones porque no es la realidad en sí misma. Por ejemplo, el aprendizaje de un grupo de estudiantes podría tener una distribución aproximadamente normal si la mayoría de los puntajes de este grupo se encuentran cercanos al promedio y unos cuantos puntajes estarían por debajo del promedio y otros cuantos por arriba de este promedio. Es decir, uno como investigador o investigadora se pregunta: ¿Cabe el modelo de la distribución normal en los datos? La distribución normal es una familia de distribuciones normales que dependen del promedio y la desviación estándar de los datos de una muestra. Por lo tanto, habría tantas distribuciones normales como promedios y desviaciones estándar fueron calculados. Existen muchas otras distribuciones de datos y habría que ver la distribución de los datos de una investigación para ver cuál sería la más adecuada.

La curva normal fue derivada de la distribución teorética binomial (Lane, Scott, Hebl, Guerra, Osherson, y Zimmer, 2014; e.g., Figura 1.1). La distribución binomial surge de la cuantificación de una serie de probabilidades de eventos donde se tiene como resultado un acierto o un fracaso: e.g., el lanzar una moneda al aire y tratar de adivinar su resultado (para más información acerca de la distribución binomial se recomienda a Kotz, 2006, y a Salkind, 2007). La curva normal puede representar varios fenómenos de la naturaleza como: la estatura o peso de las personas o su inteligencia, entre muchas otras variables. La curva normal se usa para expresar una serie de probabilidades mediante el cálculo de ciertas porciones bajo su área (Figura 1.1). En este caso, la Figura 1.1 muestra el área completa de la curva. El eje de la y representa la densidad de la probabilidad y el eje de la x representa a una variable (e.g., estatura de un grupo de personas). El área sombreada bajo la curva se puede considerar como el 100% de la probabilidad cuando se calcula del infinito negativo ($-\infty$) al positivo ($+\infty$).

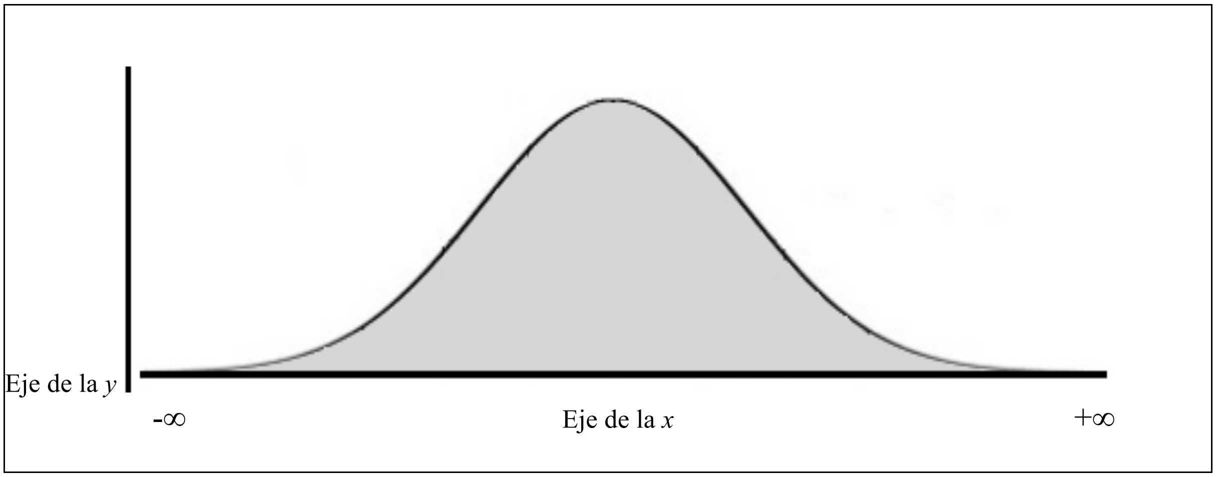


Figura 1.1 **La Curva Normal.**

La probabilidad de un número exacto es de cero porque se necesitan dos puntos para calcular el área dentro de la curva. Esta distancia entre dos puntos podría ser muy pequeña (e.g., entre 14.09 y 14.10; Figura 1.2). Sin embargo, deben de ser dos puntos diferentes para calcular un área y, con ella, una probabilidad.

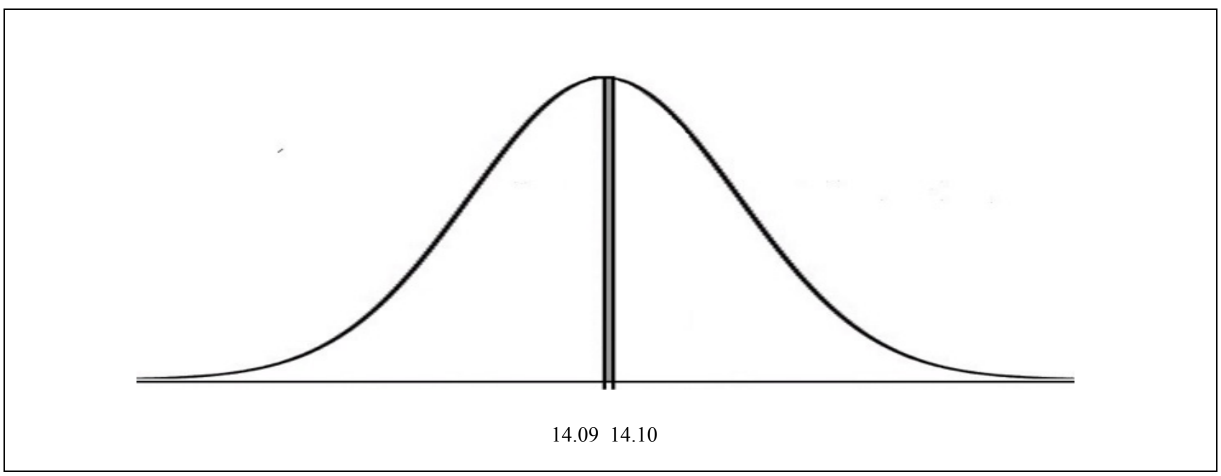


Figura 1.2 **La Curva Normal con una Pequeña Área.**

La curva normal expresa la distribución al azar de un fenómeno de una población (Hurley, 2014, p. 597), y a sus extremos también son conocidos como colas. Además, Hinkle, Wiersma, y Jurs (2003) resumieron las propiedades de la distribución normal (Tabla 1.1):

Tabla 1.1 Propiedades de la Distribución Normal.

Una distribución normal es uni-modal porque solo tiene una moda.
La variable independiente (*x*) se expresa en el eje horizontal del plano cartesiano. La variable dependiente (*y*) va en la vertical e indica una probabilidad cuando se calcula un área. En otras palabras, los valores de *y* son en función (i.e., dependen de) de los *x* (ver la Ecuación 1.1 de la curva normal más adelante).
Es simétrica porque al doblarla las dos partes son del mismo tamaño como las alas de una mariposa.
Tiene forma de campana (i.e., Campana de Gauss).
La altura máxima de la curva se alcanza en el promedio.
Es continua porque entre dos valores de x_i habrá uno más (de nuevo: e.g., entre 14.09 y 14.10 esta 14.091, entre un infinito de números más).
Es asintótica (sus dos extremos se extienden al infinito sin tocar el eje de las *x* en un plano cartesiano; desde el infinito negativo, -∞, hasta el infinito positivo, +∞).
Los valores que están 3 o más desviaciones estándar del promedio son poco probables. A estos números se les suele llamar valores atípicos (*outliers*).

La curva normal tiene la propiedad de que los valores de *y* son en función (i.e., dependen de) de los *x.* En otras palabras, si se toma un valor de *x*, se puede calcular el valor correspondiente de *y*, usando la *Formula de la Distribución Normal* (Ecuación 1.1) que se muestra a continuación:

Formula de la Distribución Normal: $y_i = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$ Ecuación (1.1)

y_i = En un plano cartesiano, es la altura de la curva para cada uno de los valores de x_i.

σ = Es la desviación estándar de una población. También, es interpretada como la desviación estándar de los datos: muestra**.**

π = Este valor de 3.1416… es una constante y representa la razón de la circunferencia al diámetro de un circulo. E.g., para la longitud de una circunferencia hay 3.1416… diámetros: 1 / 3.1416…

e = Tiene un valor constante de 2.7183… y es la base de los logaritmos naturales.

μ = Es el promedio de la población. También, es interpretada como el promedio de los datos: muestra en este tipo de casos**.**

x_i = Es valor de la variable independiente. I.e., es un punto medido por la longitud entre cero y este.

Un ejemplo con esta Formula de la Distribución Normal es: se creó un set de datos en una hoja de cálculo de Microsoft Excel que podrían representar los puntajes de un examen de admisión. Los datos que se manejaron fueron del 0 al 200 de una manera consecutiva. Es decir, se fue del 0 al 200 con una frecuencia de uno por cada valor: 0, 1, 2, 3, 4, …, 200. Se le calcularon las medidas de tendencia central: promedio ($\bar{x}$) = 100; desviación estándar (*SD*) = 58.17; mediana = 100; y como ningún valor se repitió, no hubo moda). Se aplicó la Formula de la Distribución Normal para calcular los valores de *y* con una hoja de cálculo. Con los valores de *y*, se creó la Figura 1.3. Esta Figura 1.3 también guarda algunas de las propiedades de la Tabla 1.1, excepto que no es asintótica porque tiene un rango exacto de 0 al 200: se termina en esos dos valores del rango.

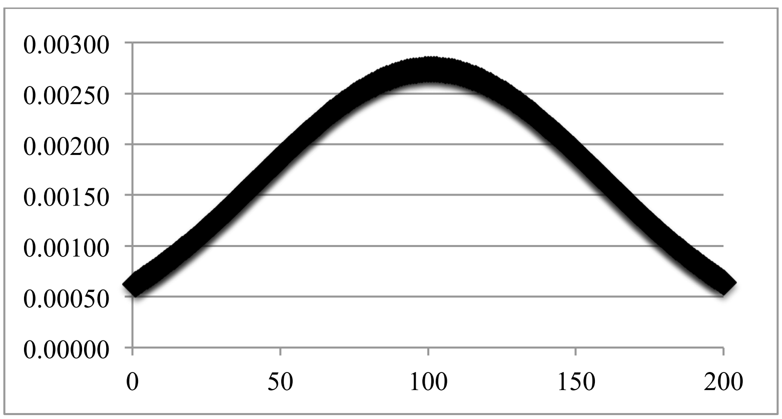


Figura 1.3 **Distribución Normal al usar la Formula de la Distribución Normal.**

a. Posible Distribución Normal de los Datos.

Antes de ahondar en las probabilidades, habría que explicar cómo se podría inferir que los datos de un investigador o investigadora parecen tener una distribución normal. Los fenómenos pueden tomar muchas distribuciones diferentes y, por ello, no van a lucir como la distribución normal. Por esta razón, es necesario cerciorarse que los datos se aproximen a una distribución normal antes de empezar a hacer análisis con ellos. Al respecto, Tabachnick y Fidell (2012, p. 79) advirtieron que “Aunque la normalidad de las variables no siempre es requerida para los análisis, la solución es usualmente mejor si todas las variables están normalmente distribuidas.”

Otra manera de poner a prueba la normalidad de los datos, con los que se va a trabajar, es usando el test de Kolmogorov-Smirnov para una muestra. Con este test, se pone a prueba la hipótesis nula de que dos distribuciones son idénticas (Salkind, 2007, p. 512). En otras palabras, se pone a prueba la hipótesis de que los datos de una muestra siguen una distribución normal. Si se rechaza la hipótesis nula, se puede concluir que la distribución de los datos es estadísticamente y significativamente diferente a la distribución normal. Este test ha sido incorporado en software como SPSS, entre otros. Esta más allá de los objetivos del presente libro explicarlo, pero Salkind (2007) hace una discusión bastante profunda sobre el tema.

b. Medidas de Tendencia Central para evaluar la Posible Distribución Normal.

Uno de los primeros pasos seria usar las estadísticas de tendencia central, i.e., promedio ($\bar{x}$), mediana y moda, para ver si tienen el mismo valor o uno semejante. Lo anterior representado matemáticamente seria:

- Cuando son iguales las tres estadísticas de tendencia central, existe evidencia de una posible distribución normal:
    - $\bar{x}$ = mediana = moda
- Cuando son semejantes las tres medidas, existe evidencia de una posible distribución normal:
    - $\bar{x} \approx$ mediana $\approx$ moda

Otro ejemplo heurístico, los datos en la Tabla 1.2 tratan de los tiempos medidos en minutos, en los cuales los estudiantes de primer grado de una primaria (n = 25) completaron un examen de matemáticas. El promedio, mediana y moda son iguales. Con estas estadísticas de tendencia central, parece que hay una distribución normal, pero aún falta por explicar los coeficientes de curtosis y asimetría que dependiendo de su valor podrían aportar evidencia para apoyar la conclusión de normalidad. La explicación de estos coeficientes se da más adelante en este capítulo del presente libro. De las medidas de dispersión, no se podría decir mucho porque no se les puede comparar por el momento con otras porque no hay otras para hacerlo.

Tabla 1.2 Datos para evaluar la Igualdad o Similitud entre Estadísticas de Tendencia Central.

Alumnos	Frecuencia	Tiempo en contestar un examen de matemáticas en minutos
Primer Grado	25	13,14,14,15,15,15,16,16,16,16,17,17,17,17,17,18,18,18,18,19,19,19,20, 20,21
	Promedio	17
	Mediana	17
	Moda	17
	Varianza	4.17
	SD	2.04
	Rango Inclusivo	(21 – 13) + 1 = 9
	Rango Exclusivo	21-13 = 8
	*Curtosis	-0.52
	*Asimetría	Cercano a 0

Nota: *Estos coeficientes son tratados más adelante: la curtosis fue -.52 y la asimetría fue cercana a 0. Al rango inclusivo abarca todos los valores desde donde parte hasta donde termina. Por esta razón, se le adiciona uno. Por el otro lado, el rango exclusivo es simplemente la diferencia entre el valor mayor y el menor.

c. Gráficas para la Normalidad para evaluar la Posible Distribución Normal.

Otra manera de detectar una distribución normal es usar graficas como una de tallos y hojas (*steam and leafs*) o una de barras. La Tabla 1.3 muestra cómo se organizaron los datos de la Tabla 1.2. El tallo, en este caso, la decena que corresponde para cada grupo de números. Es decir, el Tallo 1 significa 10 y solo habría que sumarle la unidad que corresponde y está en las hojas: e.g., Tallo 1 + Hoja 3 = 13. Las hojas también representan la frecuencia con la que aparece un número. Por ejemplo, el número cuatro aparece en dos ocasiones en la parte de las hojas, lo cual significa que el número 14 aparece dos veces en la base de datos; el 15 aparece en tres; y así sucesivamente.

Tabla 1.3 Tallos y Hojas.

Tallos	Hojas
1	3
1	4, 4
1	5, 5, 5
1	6, 6, 6, 6
1	7, 7, 7, 7, 7
1	8, 8, 8, 8
1	9, 9, 9
2	0, 0
2	1

Dada la Tabla 1.3, se podría concluir que los datos siguen aparentemente una distribución normal porque estos se distribuyen en forma de triángulo con simetría. Esta es solo una alternativa para visualizar una posible distribución normal de los datos. Otra alternativa es una gráfica de barras (Figura 1.4). En esta grafica también se forma una especie de triangulo y podría ser bordeada por una curva normal como sucede en las Figuras 1.1 y 1.2. Esta forma de la gráfica de barras también indica una posible distribución normal de los datos.

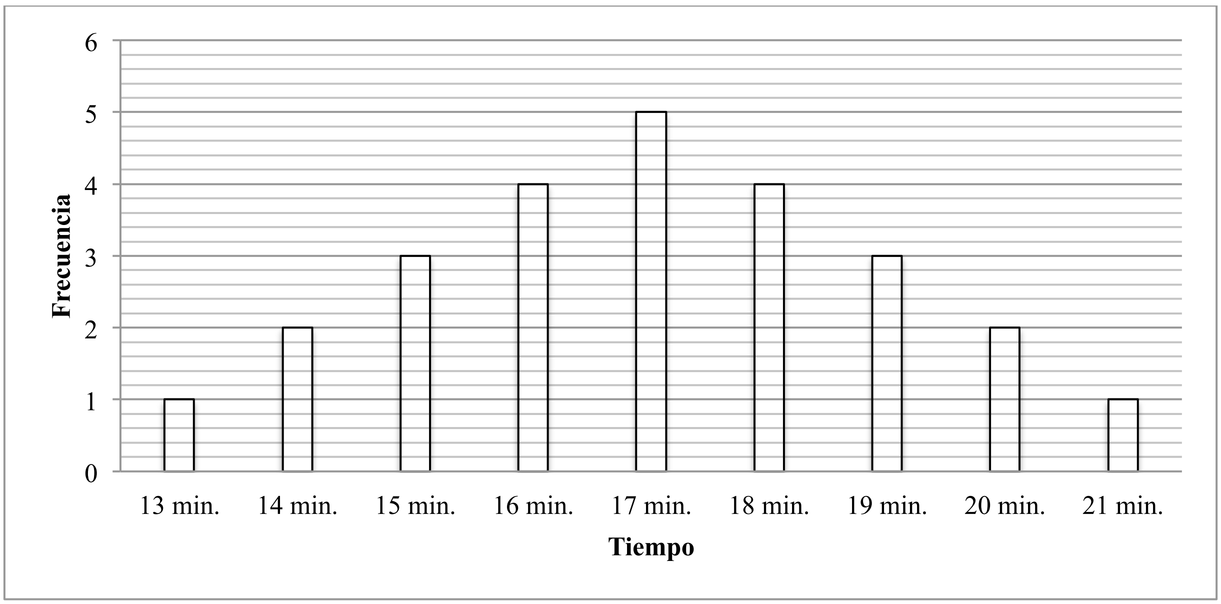


Figura 1.4 **Grafica de Barras.**

d. Curtosis y Asimetría para evaluar la Posible Distribución Normal.

Una forma diferente a las gráficas para evaluar la posible distribución normal de los datos es mediante coeficientes (curtosis y asimetría/sesgo). La curtosis (*kurtosis*) es una medición de la cantidad de datos que se concentran en las colas de una distribución con relación al resto de los datos. Cuando un set de datos es graficado, estos se distribuyen en una forma normal (la *Campana de Gauss*) con un pico en el centro y colas delgadas. La curtosis es la medida que sirve para analizar el grado de concentración de valores de una variable analizada alrededor de la zona central de la distribución de frecuencias. Más al respecto, Salkind (2011) definió la curtosis de una manera bastante elegante: La característica de una distribución que es plana o picuda. La fórmula de la curtosis (Ecuación 1.2), usada en el programa de Excel de Microsoft, es:

$$\text{Curtosis} = \left[\frac{n\,(n+1)}{(n-1)\,(n-2)\,(n-3)} \text{ X } \sum\left(\frac{(x_i - \bar{x})}{SD}\right)^4\right] - \frac{3(n-1)^2}{(n-2)\,(n-3)} \qquad \text{Ecuación (1.2)}$$

Σ = suma

n = número de observaciones (e.g., calificaciones, estudiantes, etc.)

x_i = cada uno de los valores del set

$\bar{x}$ = promedio

SD = Desviación Estándar

Nota: Esta no es la única fórmula usada en calcular la curtosis en forma manual. Por ejemplo, existen otras fórmulas para calcular la curtosis a nivel de una muestra o de una población.

Para algunos autores, estos coeficientes indican un problema cuando son calculados y su valor excede al valor absoluto de 2 o 3: i.e., |2|, |3|; recordando, las líneas paralelas indican que pueden ser valores de -2 o de +2, así como -3 o de + 3. Es decir, si son menores a -3 o mayores a +3. Asimismo, si la curtosis y la asimetría son tomadas de una muestra, son consideradas estadísticas, lo cual implica un error de medición. Si son tomadas de una población, son entonces consideradas parámetros y no tienen error de medición. Esto último es indicado porque entre las mejores prácticas que muchos autores en la investigación educativa recomiendan están los reportes de los errores de medición (para más información al respecto ver a Cumming, 2013).

Tres ejemplos de curtosis se muestran en la Figura 1.5. Cuando es normal, la curtosis es cero [Figura 1.5 (a)]. Otra posibilidad es cuando la curtosis es mayor a cero (i.e., tiene un pico), se llama leptocúrtica [*leptokurtic*, Figura 1.5 (b)]. Se convierte en un problema, según varios teóricos, cuando supera el valor de dos, en algunos casos, o de tres. La última posibilidad es cuando es menor a cero y se llama platicúrtica [*platycurtic*, Figura 1.5 (c)]. Asimismo, se convierte en un problema de curtosis una vez que supera el valor de dos o tres, según la literatura. Un ejemplo de problemas de curtosis podría pasar cuando un investigador trata de recolectar datos a través de una encuesta. Un problema de distribución leptocúrtica seria cuando en esta encuesta de cinco puntos la mayoría de los participantes seleccionara el valor tres de la escala. De este modo, la mayoría de los datos se concentrarían en el centro de una distribución, trayendo consigo un posible coeficiente de curtosis > 0 y un problema de leptocúrtica. Por otro lado, podría pasar que los participantes distribuyeran sus respuestas a lo largo de la escala de cinco puntos. De este modo, las colas de la distribución tendrían más datos y el coeficiente de curtosis < 0. Esto podría traer un problema de una distribución platicúrtica.

Salkind (2007, p. 522) presento una serie de procedimientos para calcular la significancia estadística de un coeficiente de curtosis. En otras palabras, el coeficiente de curtosis puede ser considerado como un problema dependiendo del grado del alfa que se seleccione para el test de significancia (e.g., alfa = .05, .01, y .001) y el tamaño de la muestra de donde se obtuvo.

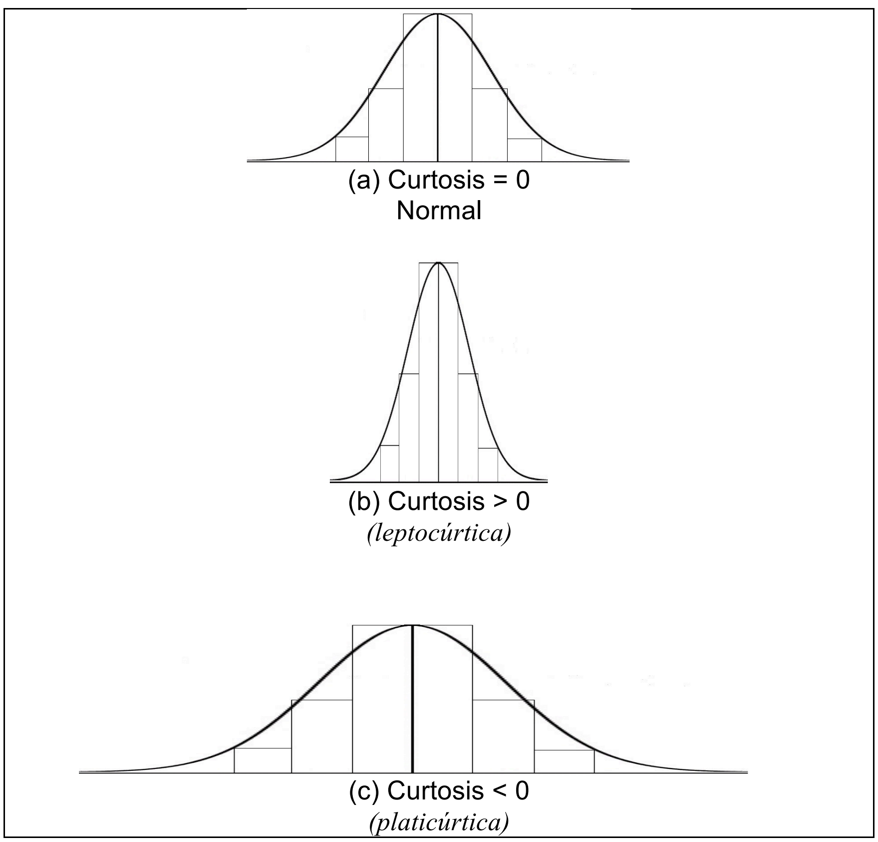


Figura 1.5 **Ejemplos de Curtosis.**

Asimetría (*Skewness*) también es conocida como sesgo. Según Salkind (2011, p. 435), es una característica de una distribución que muestra una frecuencia desproporcionada de ciertos datos. Mide el grado de asimetría de un set de datos. La fórmula de la asimetría (Ecuación 1.3), usada en el programa de Excel de Microsoft, es:

$$\text{Asimetría} = \frac{n}{(n-1)(n-2)} \; X \; \Sigma \left(\frac{x_i - \bar{x}}{SD}\right)^3 \qquad \text{Ecuación (1.3)}$$

Σ = suma

n = número de observaciones (e.g., calificaciones, estudiantes, etc.)

x_i = cada uno de los valores del set

$\bar{x}$ = promedio

SD = Desviación Estándar

Nota: Esta no es la única fórmula usada en calcular la asimetría en forma manual. Por ejemplo, existen otras fórmulas para calcular la asimetría a nivel de una muestra o de una población.

En la Figura 1.6, se muestra una serie de ejemplos de asimetría. El ejemplo (a) muestra el caso en el que la asimetría es cero, lo cual indica que la curva es perfectamente simétrica. Por otro lado, la distribución de la (b) esta sesgada positivamente porque el promedio es mayor que la mediana y esta última es mayor que la moda. Por lo tanto, la asimetría es mayor que cero. Esto ocasiona que la cola del lado derecho sea más alargada que la otra.

Esto puede pasar en la investigación educativa cuando uno o más valores sean muy superiores al resto. Por ejemplo, una investigadora está recolectando datos de los estudiantes para estimar cuanto es su gasto mensual (e.g., renta, transporte, diversión, vestido, comida, etc.). La pregunta de investigación descriptiva puede ser: ¿Cuál es el gasto mensual de los estudiantes universitarios? Toma una muestra de 100 participantes que ella considera representativa y, bajo un contrato, les pregunta en forma confidencial el monto en pesos que gastan en un mes típico. Los datos indican que el rango va aproximadamente desde 3,000 pesos a 5,000. Sin todavía haber calculado estadísticas de tendencia central o dispersión, se da cuenta que uno de los estudiantes reporto gastar aproximadamente 200,000 pesos al mes. Como pudo haber sido un error de captura de datos, le vuelve a preguntar a este estudiante quien confirma que gasta estos 200,000 al mes. La razón de tener estos gastos es que está comprando, con una parte de una herencia recientemente adquirida, obras de arte que algún día piensa vender. La investigadora calcula las estadísticas y, por la simplicidad del argumento, se usan estadísticas aproximadas para ilustrar este ejemplo del gasto:

$\bar{x}$ = 6,000

Mediana = 4,000

Moda = 3,900

SD = 20,000

Asimetría = 10

Ella se da cuenta que el coeficiente de asimetría es 10 que es varias veces mayor al que se considera el limite |3|. Para solucionar este, problema decide eliminar el gasto de 200,000. Ahora, sus estadísticas son las siguientes:

$\bar{x}$ = 4,000

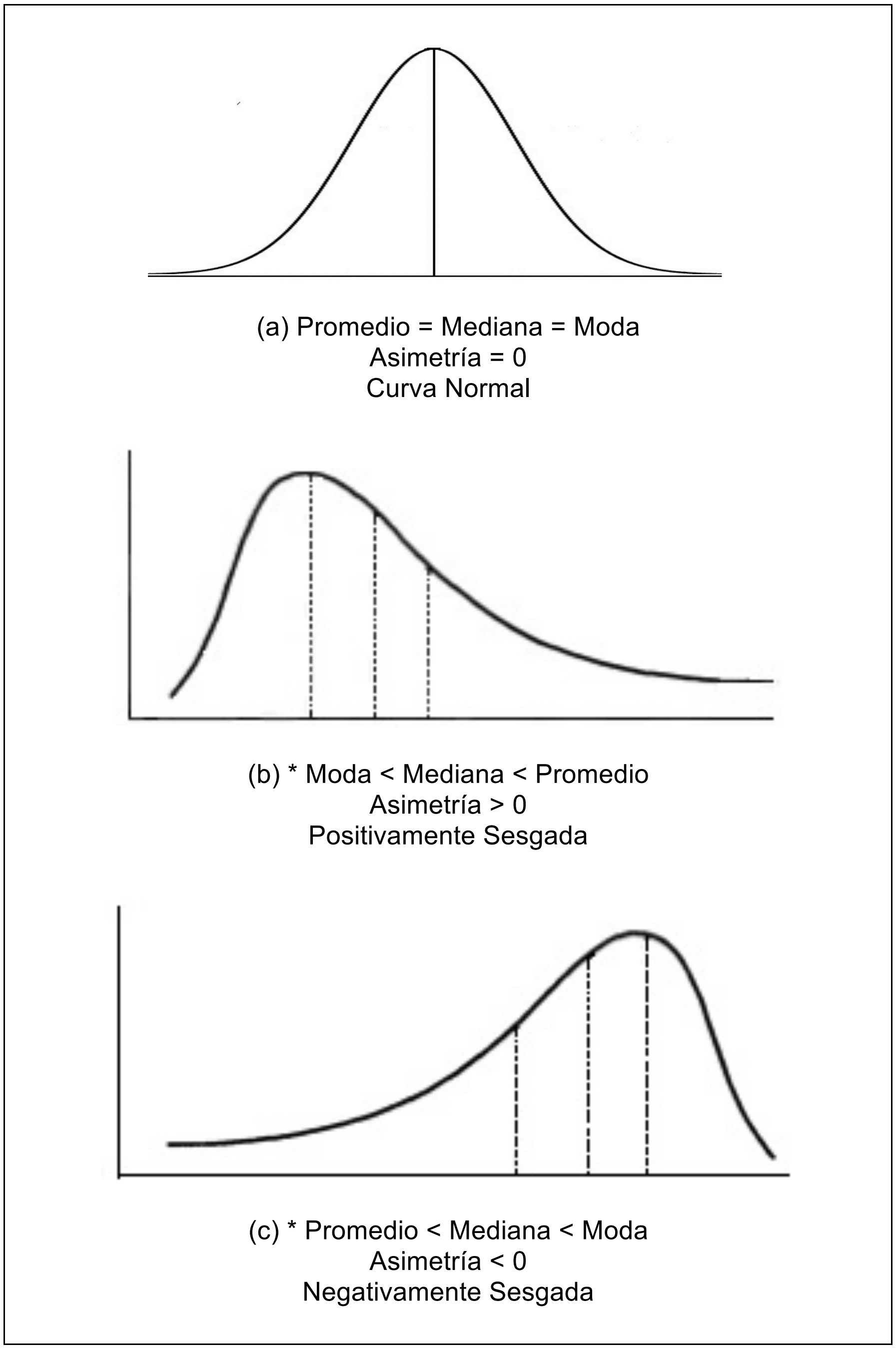
Mediana = 4,000

Moda = 3,900

SD = 1

Asimetría = 0.50

El resultado de la eliminación de los datos del estudiante antes mencionado fue que las estadísticas de tendencia central son casi todas iguales. Asimismo, el coeficiente de asimetría cayó a un nivel aceptable. Para una discusión más profunda sobre como estimar cuando la asimetría es problemática, se recomienda a Salkind (2007, p. 918). En conclusión, la investigadora concluye que sus datos aparentemente tienen una distribución normal porque las estadísticas de tendencia central son similares y el coeficiente de asimetría es menor que tres.



Nota: *De izquierda a derecha, la línea punteada representa el orden de la estadística.

***Figura 1.6* Ejemplos de Asimetría.**

En caso de tener algún problema de curtosis o asimetría, Tabachnick y Fidell (2012) han recomendado la transformación de datos. Una transformación implica usar algún exponencial o logaritmo para que los coeficientes de la curtosis y asimetría alcancen. Por ejemplo, si se tiene un set de valores como 3, 4, y 7 bajo el supuesto de que hay una asimetría > 3, se puede elevar al cuadrado cada valor: 2^2, 4^2, y 7^2. Queda 4, 16, y

49 que son los datos transformados y, de donde, se recalcularía la curtosis y asimetría para ver si se acercaron a 0. Para más información sobre transformaciones, ver a Tabachnick y Fidell (2012).

Preguntas para Reflexionar

1.- ¿Qué otras distribuciones se podrían usar para calcular las probabilidades de ciertos datos?

2.- ¿Qué consecuencias puede traer el usar los datos que no sigan una distribución normal?

3.- ¿Cómo se podrían corregir los problemas de asimetría?

4.- ¿Cómo se podrían corregir los problemas de curtosis?

Opinión del Autor

Creo que lo más importante de esta sección es darse cuenta que el modelo de la distribución normal puede ser utilizado para la investigación educativa. Esto se puede hacer cuando nuestros datos tienen ciertas características de medidas de tendencia central similares, cierta dispersión de los datos y coeficientes aceptables de curtosis y simetría. Dada la situación de que los coeficientes de los datos, hay maneras para corregir los datos.

En mi experiencia como profesor, he observado que los estudiantes de educación llevan la clase de estadísticas antes que las clases de sus proyectos de tesis. Sin embargo, la mayoría no hace la conexión entre lo que aprendieron en la clase de estadísticas y su posible uso en su proyecto de investigación. Lo que suelo hacer es repasar los principios, análisis, graficas, etc. de estadísticas usando ejemplos educativos. Entonces, me ha pasado que si logran hacer la conexión.

Capítulo 2. La Distribución Normal Estandarizada.

La distribución normal estandarizada podría ser definida como la manera de poner en practica la distribución normal en algún tipo de investigación. Este modelo de la distribución normal estandarizada puede brindar ahorros de tiempo y esfuerzo para calcular probabilidades. De otra manera, un investigador o investigadora tendría que usar calculo integral para calcular las áreas bajo la curva que representan las probabilidades.

Hinkle et al. (2003) explicaron que para poder utilizar el modelo de la distribución normal a los datos de una investigación es necesario usar la *distribución normal estandarizada*. Esto se debe a que habría una infinidad de distribuciones normales debido a que cada promedio y desviación estándar las producirían. Además, al usar la *distribución normal estandarizada* se pueden utilizar las tablas de los valores z que aparecen en muchos libros de estadísticas para calcular las probabilidades de los valores (ver Apéndice). Un valor z se obtiene mediante la siguiente formula (Ecuación 2.1):

$$z_i = \frac{(x_i - \bar{x})}{SD} \qquad \text{Ecuación (2.1)}$$

z_i = es el coeficiente que se obtiene de la operación y puede ser localizado en la *distribución normal estandarizada.*

x_i = es el valor conocido como crudo o bruto de la recolección de datos. Se encuentra en la unidad en la que se recolectaron los datos: e.g., cm, kilos, etc.

$\bar{x}$ = es el promedio de los datos crudos o brutos en su unidad.

SD = es la desviación estándar de los datos crudos o brutos.

En otras palabras, si se tiene un valor individual en bruto, el promedio de los datos en bruto y su *SD* en bruto, se puede calcular su correspondiente valor estandarizado (z_i). Con este valor z, se puede saber su ubicación en la *distribución normal estandarizada* (Figura 2.1). Los valores (|1|, |2|, y |3|) que se muestran en la Figura

2.1 también coinciden con la distancia en desviaciones estándar con el promedio. Ejemplos, un valor $z = |1|$ es igual a una $SD = |1|$; $z = |2|$ igual a dos $SD = |2|$; y $z = |3|$ igual a tres $SD = |3|$.

Cada punto en la línea de *distribución normal* estandarizada puede ser calculado de la misma manera que los puntos en la curva normal. Similarmente, la *distribución normal* estandarizada tiene la propiedad de que los valores de y son en función (i.e., dependen de) de los x. En otras palabras, si se toma un valor de x, se puede calcular el valor correspondiente de y, usando la *distribución normal* estandarizada (Ecuación 2.2) que se muestra a continuación:

$$y_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i{}^2}{2}} \qquad \text{Ecuación (2.2)}$$

y_i = En un plano cartesiano, es la altura de la curva para cada uno de los valores de x_i.

π = Este valor de 3.1416 es una constante y representa la razón de la circunferencia al diámetro de un circulo. E.g., para la longitud de una circunferencia hay 3.1416… diámetros: 1 / 3.1416…

e = Tiene un valor constante de 2.7183… y es la base de los logaritmos naturales.

z_i = Es valor de la variable independiente. I.e., es la longitud.

El área bajo la distribución normal estandarizada es del 100%. También, esta área se expresa como la probabilidad de toda el área es uno.

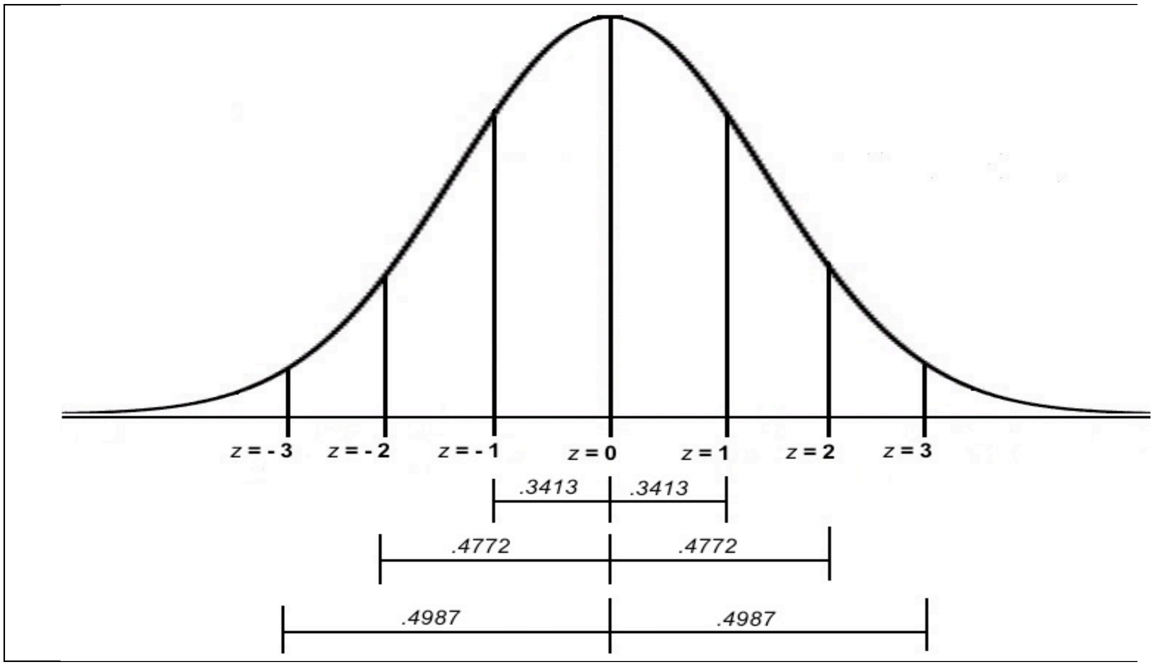


Figura 2.1 **Áreas en la *Distribución Normal Estandarizada.***

a. Un Ejemplo de la Distribución Normal Estandariza.

Un ejemplo de los valores *z* seria usando valores en bruto de los puntajes de coeficiente intelectual de un examen como el Test de Inteligencia de Wechsler (Wechsler, 1997). El puntaje total de este test contiene los puntajes de las diferentes áreas de la inteligencia que son medidas (e.g., lenguaje, matemáticas, espacio, etc.). El puntaje promedio es de 100 puntos con desviación estándar de 15 puntos. En otras palabras, un o una estudiante que tome este test tendrá una probabilidad del 68.26% (Tabla 2.1; 68.29%) con un puntaje de 85 a 115 puntos. Sería poco probable (i.e., 0.26%) que este o esta estudiante obtuviera un puntaje de menos de 55 o más de 145 puntos.

Tabla 2.1 Probabilidad, Valores *z*, y Puntajes del Wechsler.

Valores *z* / Desviación Estándar	Área / Probabilidad	Puntaje en el Wechsler	Probabilidad en % de que un estudiante este en este rango.
De -1 a 1	.3413 + .3413 = .6826	85 a 115 puntos	68.26
De -2 a 2	.4772 +.4772 = .9544	70 a 130 puntos	95.44
De -3 a 3	.4987 + .4987 = .9974	55 a 145 puntos	99.74
Más allá de -3 y 3	.0026	Menos de 55 o más de 145 puntos	0.26

b. Usos de la Distribución Normal Estandarizada.

En la investigación educativa, el uso de los valores *z* puede ayudar a ubicar a los participantes de algún test o encuesta en la distribución normal estandarizada. Al poderlos ubicar en esta curva, se pueden calcular diferentes probabilidades, dependiendo del valor *z*. En el ejemplo anterior del Test de Wechsler, se muestra esta relación entre el rango de área que se cubre entre el promedio ($z = 0$) y un valor *z* (i.e., |1|, |2|, y |3|).

Para ilustrar la ubicación de participantes en la distribución normal estandarizada se usan los datos de Ponce, Soto y Solorio (2017). Estos autores compararon los puntajes de diferentes grupos de aspirantes a una universidad del norte de México en el examen de admisión: Examen de Habilidades y Conocimientos Básicos (EXHCOBA). Las estadísticas e información del estudio de los autores antes mencionados fueron las siguientes: los aspirantes tomaron el examen en el otoño del 2016 ($n = 4{,}785$) para iniciar el semestre en enero del 2017; treinta y cuatro programas de pregrados/carreras habían sido ofertados; el promedio de todos los aspirantes fue de 78.61 puntos y su desviación estándar = 27.89 puntos; y solo se tomaron tres carreras para este ejemplo (i.e., medicina, ingeniería en mecatrónica y psicología).

La fórmula para obtener un valor *z* individual es la siguiente: $z_i = (x_i - \bar{x}) / SD$. Sin embargo, cuando se usan muestras en lugar de individuos es más apropiado el uso de la siguiente formula: $z_i = (x_i - \bar{x}) / (SD / \sqrt{n})$. Para más información al respecto del uso de fórmulas para muestras o individuos, se recomienda ver a Johnson (1990). Por simplicidad para este ejemplo de los aspirantes a estas tres carreras, se usa el promedio sus puntajes como un valor *z* individual. Entonces, la Ecuación 2.3 para el ejemplo es:

$$z_{\,carrera} = \frac{(x_{\text{promedio del puntaje de la carrera}} - \bar{x}_{\text{promedio del puntaje de todos los aspirantes}})}{SD_{\text{puntaje de todos los aspirantes}}} \qquad \text{Ecuación (2.3)}$$

La Tabla 2.2 muestra los datos de estas carreas. También, se aplica la fórmula de los valores *z* para el cálculo de estos valores por carrera.

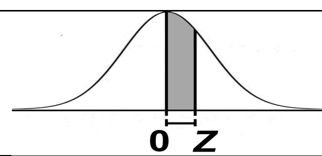
Tabla 2.2 Calculo de los Valores *z* por Carrera.

Carrera	*n*	Promedio de Puntos en el EXHCOBA por carrera	Valor *z*	Distancia entre el promedio y probabilidad.
medicina	731	105.70	$z_{\text{medicina}} = (105.70 – 78.61) / 27.89 = 0.97$	$z = 0.97$ 33.40%
ingeniería en mecatrónica (IM)	138	78.82	$z_{\text{IM}} = (78.82 – 78.61) / 27.89 =$ 0.007 redondeado a la centésima más cerca 0.01	$z = 0.01$ 0.40%
psicología	267	65.10	$z_{\text{psicología}}$ $(65.10 – 78.61) / 27.89 = - 0.48$	$z = - 0.48$ 18.44%

Nota: El promedio de todos los aspirantes fue de 78.61 y *SD* = 27.89.

El primero es el de los aspirantes a la carrera de medicina ($z = 0.97$). En el Apéndice de la Tabla de los Valores *z* para la distribución normal estandarizada, se indica que la probabilidad del promedio ($z = 0$) y el valor *z* de medicina (0.97) es de .3340. La Tabla 2.3 muestra una porción de la Tabla de los Valores *z* para la distribución normal estandarizada. El procedimiento para identificar la probabilidad asociada con cierto valor *z*, es ubicar el valor *z* en la columna que dice: Valor z_i. En seguida, la probabilidad entre este valor *z* y el promedio ($z = 0$) aparece al costado derecho (e.g., Tabla 2.3).

Tabla 2.3 Porción de la Tabla de los Valores *z*.

	
Valor z_i	Área entre el promedio y z_i
0.97	.3340

Es decir, el área del valor $z = 0$ y $z_{medicina}$ de 0.97 es del 33.40%. La Figura 2.2 indica esta área del 33.40%. Esto puede interpretado como que la muestra de aspirantes tuvo la probabilidad (33.40%) de tener un valor *z* entre 0 y 0.97.

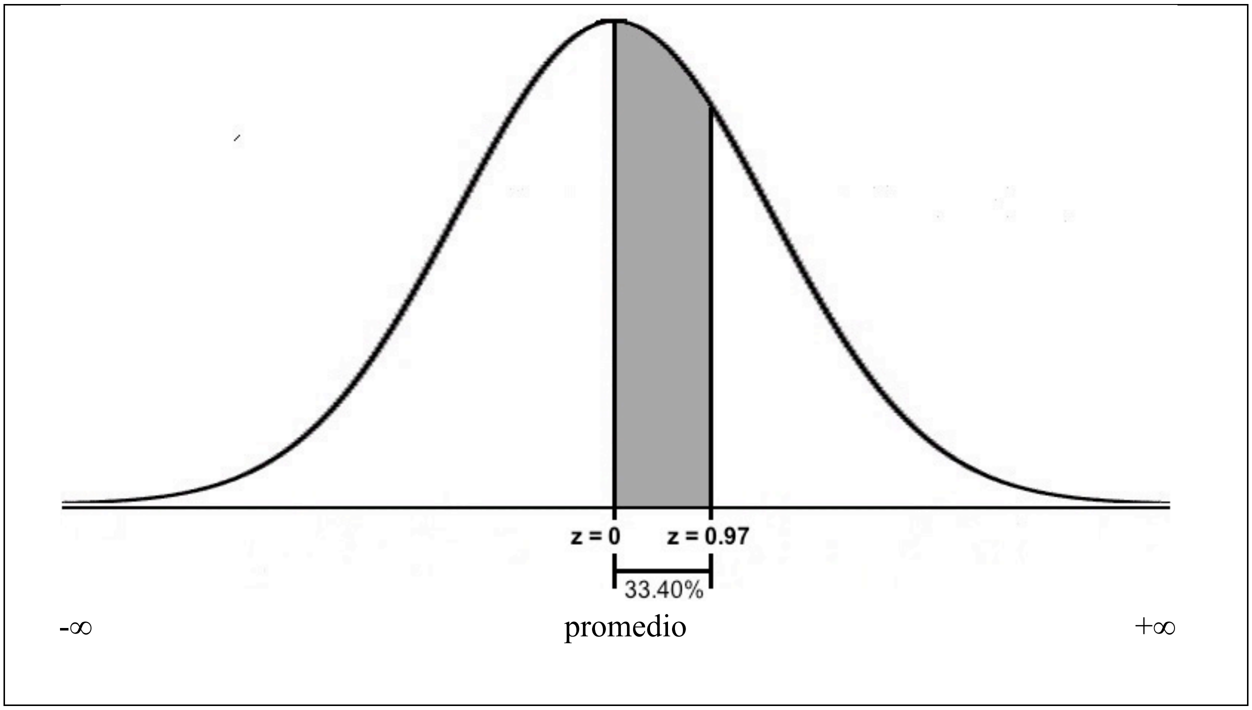


Figura 2.2 **Valor *z* de los Aspirantes de Medicina.**

Otra interpretación a la que se puede llegar es que la probabilidad de que alguien hubiera obtenido un puntaje superior al de medicina fue del 16.60% (Figura 2.3; entre los valores z de 0.97 y el infinito positivo). Recordando, la probabilidad es del 50% de obtener un valor z de 0 al infinito ($|\infty|$). La siguiente operación muestra lo anterior:

Área entre $z = 0$ y $z = 0.97$ es de 33.40%.

Área entre $z = 0$ y el $+\infty$ es de 50%.

La diferencia entre ambas áreas da el área restante: 50 – 33.40 = 16.60%

Otra operación que se puede llevar a cabo es la probabilidad de haber obtenido un promedio de los puntajes en el examen de admisión menor al valor z de 0:

Área entre $z = 0$ y el $-\infty$ es de 50%.

Una interpretación más a manera de pregunta: ¿Cuál fue la probabilidad de estar por abajo del promedio del puntaje de todos los aspirantes y por encima de los puntajes de los de medicina? La respuesta seria la siguiente:

La diferencia entre ambas áreas da el área restante: 50 ($z = 0$ y el $+\infty$) – 33.40 ($z = 0$ y $z = 0.97$) = 16.60%

Área entre $z = 0$ y el $-\infty$ es de 50%.

Sumando ambas áreas: 16.60 + 50 = 66.60%

La respuesta a la pregunta anterior es que existe una probabilidad de 66.60% de estar por debajo del promedio de los aspirantes y por encima del promedio de puntos de medicina.

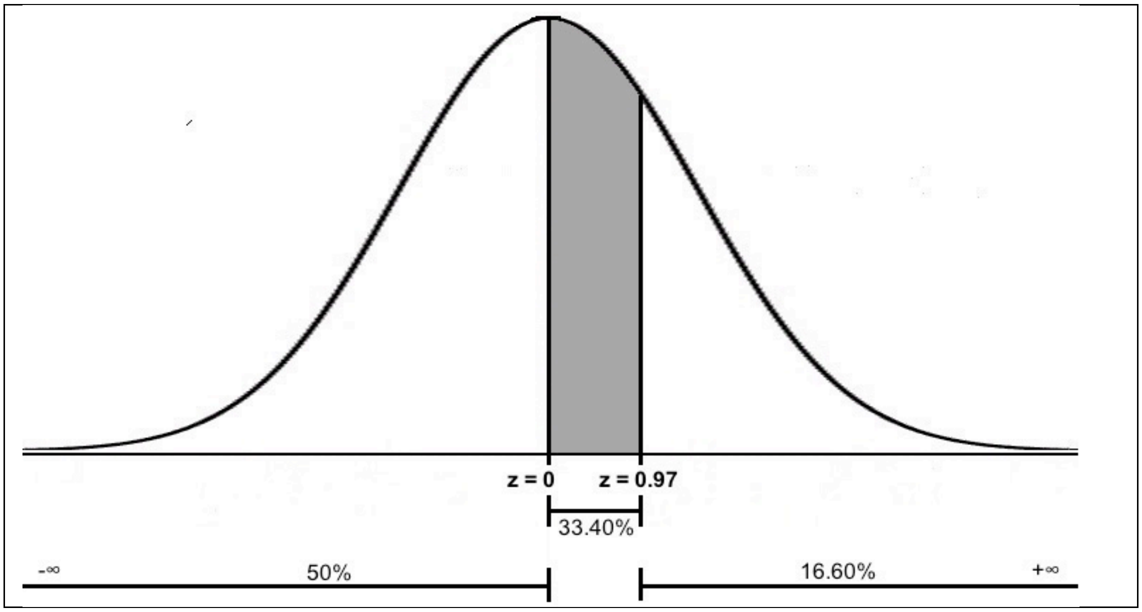


Figura 2.3 **Valor *z* de los Aspirantes de Medicina y otras Áreas de la Distribución Normal Estandarizada.**

Los aspirantes a la carrera de ingeniería mecatrónica obtuvieron un valor z de 0.007, pero fue redondeado a 0.01 (Tabla 2.2), por simplicidad para elaborar el ejemplo. Una interpretación de las probabilidades de los aspirantes a ingeniería mecatrónica es las siguiente (Figura 2.4). El área entre el valor $z = 0$ y el $z = 0.01$ es de .004 (Apéndice; Tabla de Valores z). Esto indica que la probabilidad fue del 0.40% de haber estado entre el promedio de todos los aspirantes a los 34 pregrados y el puntaje promedio de los aspirantes de ingeniería mecatrónica. Una pregunta que puede surgir: ¿Cuál fue la probabilidad de estar por encima del promedio del puntaje de los aspirantes de ingeniería mecatrónica?

Área entre $z = 0$ y $z = 0.01$ es de .04%.

Área entre $z = 0$ y el $+\infty$ es de 50%.

La diferencia entre ambas áreas da el área restante: 50 – .04 = 49.60% (respuesta a esta última pregunta: un 49.60% estaría por encima del promedio y de los aspirantes a la carrera de ingeniería mecatrónica).

Otra posible pregunta es: ¿Cuál fue la probabilidad de estar por abajo del promedio del puntaje de todos los aspirantes y por encima de los puntajes de los aspirantes a ingeniería mecatrónica? La respuesta seria la siguiente:

La diferencia entre ambas áreas da el área restante: 50 ($z = 0$ y el $+\infty$) – 0.4 ($z = 0$ y $z = 0.01$) = 49.60%

Área entre $z = 0$ y el $-\infty$ es de 50%.

Sumando ambas áreas: 49.60 + 50 = 99.60% (la respuesta a esta última pregunta es que casi toda la muestra de aspirantes estaría por abajo y por encima del promedio de los aspirantes de mecatrónica). En otras palabras, fue casi el 100% de probabilidades de que los puntajes de los 33 programas restantes estuvieran por debajo o por encima de los puntajes promedio de los aspirantes al pregrado de ingeniería mecatrónica.

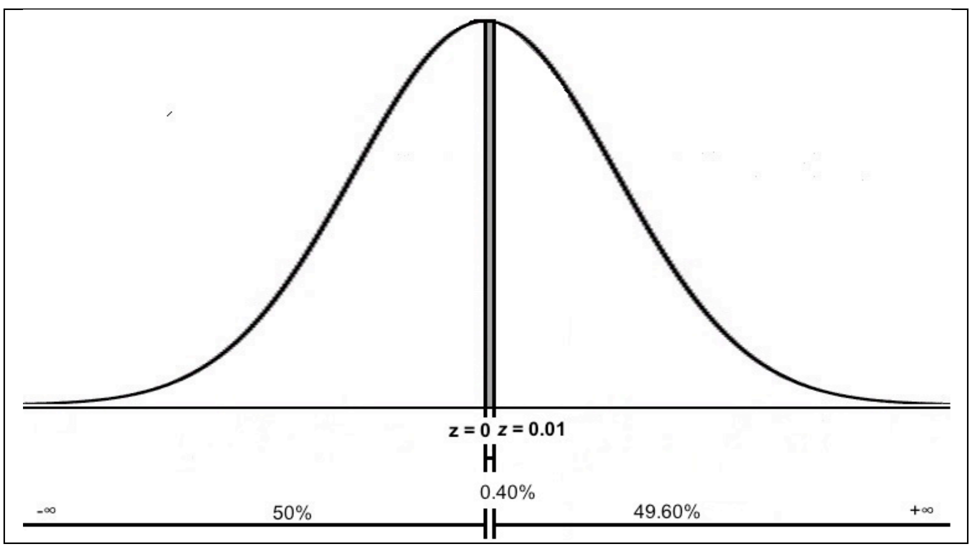


***Figura 2.4* Valor *z* de los Aspirantes de Ingeniería Mecatrónica y otras Áreas de la Distribución Normal Estandarizada.**

Una interpretación de las probabilidades de los aspirantes a psicología es la siguiente (Figura 2.5). El área entre el valor $z = 0$ y el $z = -0.48$ es de .1844 (Apéndice; Tabla de Valores z). Aunque el valor z tiene un signo negativo, no significa que la probabilidad va a ser negativa. De hecho, no existe una probabilidad negativa. Este signo negativo solo indica el sentido para ubicar el valor z calculado. Los valores z positivos van a la derecha del valor $z = 0$, y los negativos a la izquierda de este. Volviendo al cálculo del área, la probabilidad fue del 18.44% de haber estado entre el promedio de todos los aspirantes a los 34 pregrados y el puntaje promedio de los aspirantes de psicología. Una pregunta que puede surgir: ¿Cuál fue la probabilidad de estar por encima del promedio del puntaje de los aspirantes de psicología?

Área entre $z = 0$ y el $+\infty$ es de 50%.

También, otra pregunta es: ¿Cuál es el área por debajo del valor z de los aspirantes de psicología?:

La diferencia entre ambas áreas da el área restante: 50 ($z = 0$ y el $-\infty$) – 18.44 ($z = 0$ y $z = -0.48$) = 31.56%.

Otra posible pregunta es: ¿Cuál fue la probabilidad de estar por abajo del promedio del puntaje de los aspirantes de psicología y por encima de los puntajes de todos los aspirantes? La respuesta seria la siguiente:

Área entre $z = 0$ y el $+\infty$ es de 50%.

Sumando ambas áreas: 31.56 + 50 = 81.56% (la respuesta a la pregunta es que un porcentaje que se le podría considerar grande estaría por debajo o por encima de del promedio de los aspirantes de psicología).

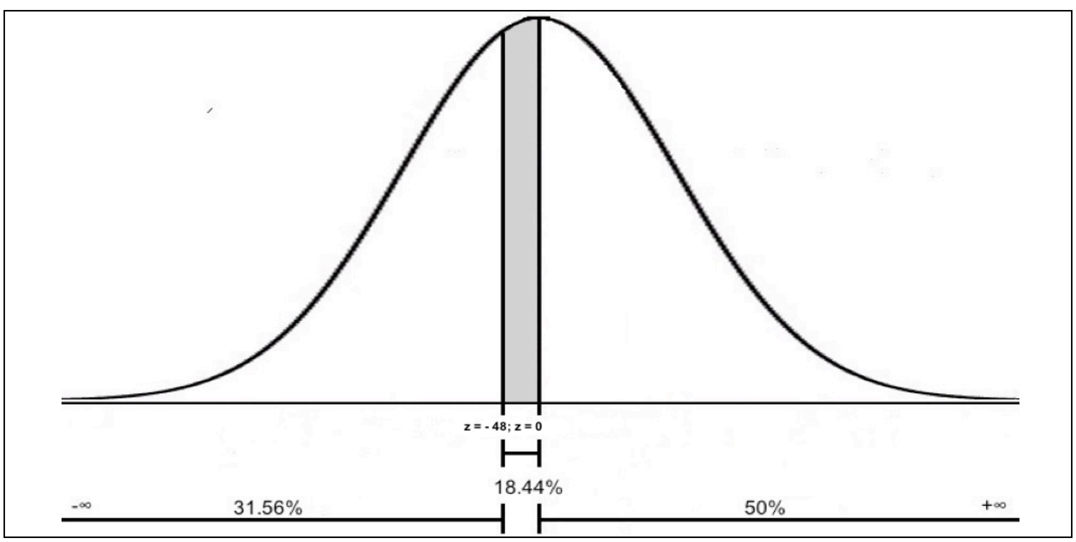


Figura 2.5 **Valor *z* de los Aspirantes de Psicología y otras Áreas de la Distribución Normal Estandarizada.**

En resumen, los valores z de los aspirantes a entrar a los tres pregrados antes descritos fueron mayores a $z = -1$ y menores $z = +1$. Estos valores estuvieron dentro del área de mayor probabilidad de la distribución normal estandarizada: 68.26% (ver Figura 2.1 para ubicar estos valores z). De nueva cuenta, estas probabilidades son aproximadas cuando los datos siguen hasta cierto punto una distribución normal.

Entonces, es factible utilizar la distribución normal estandariza. El uso de esta distribución economiza el cálculo de áreas bajo una curva a través del cálculo integral.

Preguntas para Reflexionar

1.- ¿Se pueden estandarizar todos los valores de un set de datos?

2.- ¿De qué otras maneras se podrían organizar los datos para estimar sus probabilidades?

3.- ¿En qué tipo de investigación educativa serían más útiles los valores z?

4.- ¿Qué se podría hacer si los datos aparentemente no siguen una distribución normal?

5.- ¿Cómo se podría sacar el área bajo la curva sin utilizar la tabla de los valores z?

Opinión del Autor

Más que la opinión personal, es un pequeño resumen de esta sección. La distribución normal estandariza es un modelo que se usa en lugar de la distribución normal porque esta última puede variar su forma dependiendo del valor del promedio y de la desviación estándar. Lo que resulta de la variación en los promedios y desviación estándar es una familia de distribuciones normales. Sin embargo, esta familia de distribuciones no resulta practica a la hora de calcular las probabilidades porque involucra el uso del cálculo integral para calcular un área dentro de cada una de las distribuciones. Por estas razones y bajo el supuesto de que los datos siguen más o menos una distribución normal, se usa la distribución normal estandarizada que no varía cuando un promedio o una desviación estándar de una muestra en particular cambian.

En investigación educativa, podemos transformar los datos en forma individual o por grupos en valores *z*. Con estos valores, podemos saber que tan lejos del promedio están y en qué dirección: por arriba o por abajo del promedio. Esta información sirve para consultar una tabla de valores *z* y calcular la probabilidad de una o varias áreas. Estas probabilidades se interpretan de acuerdo con la pregunta de investigación que estemos siguiendo.

Capítulo 3. Estadísticas Inferenciales.

Las personas tienden a generalizar acerca de muchas cosas: e.g., otras personas, lugares, etc. Unos ejemplos de esto son dichos, tales como: "Todos los hombres son iguales" o "Para muestra, basta un botón." Otros ejemplos serian que tal religión es de fanáticos o cierto grupo étnico está constituido por ladrones. Muchas veces, se perciben a los demás de acuerdo con una o varias experiencias que se han tenido. A veces, las experiencias ni siquiera son propias porque se escucharon de alguien más, y así una persona se formó su propia idea de algo o alguien. Algunos psicólogos llaman a esto último el "ser prisioneros de nuestra propia experiencia." Para tratar de escapar esta prisión hecha de experiencias y, en muchas de las ocasiones, de anécdotas de otras personas, se puede llevar a cabo una investigación donde se puedan hacer inferencias estadísticas (generalizaciones). Para generalizar en la investigación educativa, es necesario utilizar muestras y estadísticas para poder concluir que posiblemente todos los hombres de una población (varones estudiantes, padres o maestros) tienen cierto atributo, siguiendo un poco al dicho antes mencionado: "todos son iguales." Al usar muestras, hay imprecisiones así que sería más apropiado decir: "Parece que casi todos son iguales."

a. Resumen de la Curva Normal.

Por un lado y para Fischer (2010, p. 75), el objetivo principal de la *Teoría Clásica de la Probabilidad* fue en sus principios usada para calcular las probabilidades de ciertos eventos, con la meta de hacer decisiones racionales basadas en estas probabilidades. Por otro lado, y según Ross (1997, p. 257), la *Teoría Clásica de la Probabilidad* (*Probability Theory*) es una rama de las matemáticas que juega un papel central en las estadísticas *inferenciales*. Este último autor explicó que este papel central se debe a dos procedimientos, en los cuales se usa esta *teoría*:

I.- Estimación de *parámetros*: Es inferir de las estadísticas de las muestras (e.g., promedio = $\bar{x}$) los parámetros de una población (e.g., promedio = μ). Este concepto se explica más en detalle en los próximos párrafos.

II.- Poner a prueba *hipótesis*: Es usar un test para poder inferir si la diferencia entre dos promedios fue probablemente al azar. Este tema se tratará más delante bajo el nombre de *Test de Significancia Estadística de la Hipótesis Nula* (i.e., *Null Hypotesis Significanse Testing*).

Con la Teoría Clásica de la Probabilidad, se pueden elaborar modelos para explicar algún fenómeno. Los modelos estadísticos pueden venir de teorías que especifican relaciones entre variables (e.g., causa y efecto). También, los modelos pueden ser creados de los datos para explicar relaciones entre variables independientes con una dependiente como se hace en algunos casos con las regresiones (para los modelos de regresión, ver a Tabachnick y Fidell 2012). También, este tema de explorar las relaciones y diferencias entre variables se logra mediante el uso de minería de datos (ElAtia, Ipperciel, y Zaiane, 2016). Los modelos toman forma de ecuaciones o funciones matemáticas, y el siguiente es un ejemplo de esto:

y = 2x

y = calificaciones en un parcial de matemáticas.

x = horas de estudio para el primero de los parciales de matemáticas.

En palabras, la calificación depende de la cantidad de horas de estudio multiplicadas por dos. Este modelo será explicado más adelante en la sección de *Variables Dependientes e Independientes*. Por el momento, solo es importante mostrar cómo es un modelo y que variables podría contener.

"Cualquiera modelo realista de un fenómeno del mundo real *debe* de tomar en cuenta la posibilidad de aleatoriedad" (Ross, 1997, p. 1). Este último autor explico al respecto que al medir un fenómeno este tendrá inherentemente variación que debe de ser considerada en el modelo, y, por ello, los modelos son probabilísticos en su naturaleza. Un modelo puede ser la curva normal que puede representar la distribución de una serie de fenómenos educativos (e.g., aprendizaje, inteligencia, autoestima, etc.). Es decir, un promedio de uno de estos fenómenos (μ) debajo de la *curva normal* es el valor más probable de una población. En otras palabras, al medir el aprendizaje, inteligencia o autoestimada de una población, los promedios de estos constructos serán los más probables. Sin embargo, hay que tomar en cuenta esta variación asociada con el azar (esto se desarrollara a través del libro).

Las estadísticas inferenciales permiten a los investigadores e investigadoras hacer generalizaciones acerca de que tan bien las estadísticas de las muestras corresponden a los parámetros de las poblaciones (i.e., Estimación de parámetros; Salkind, 2007, p. 457). En una curva normal, se representa a una variable x (e.g., calificaciones, inteligencia, autoestima, etc.). La forma de la distribución normal (N) depende del valor del promedio de la población (μ) y de la varianza de la población (σ^2). Lo anterior se puede representar de esta manera:

N (μ, σ^2)

N = distribución normal de una población.

μ = promedio de la población.

σ^2 = varianza de la población.

De nuevo, la forma de la curva normal está en función de los valores del promedio y la varianza. Otra fórmula seria con la substitución de la varianza (σ^2) por la desviación estándar (σ). La nomenclatura anterior es cuando se habla de poblaciones. Cuando se habla de muestras, la nomenclatura cambia a N ($\bar{x}$, varianza) o expresado con la desviación estándar: N ($\bar{x}$, *SD*). La variable x aparece al azar (i.e., cuando se toma una muestra aleatoriamente), tiene cierto rango y cada uno de sus valores tiene cierta probabilidad. Aparecer al azar significa que una variable como un nivel de aprendizaje en una población surge al tomar una muestra al azar de una población. Por ejemplo, de una población se toma una muestra al azar para estimar su promedio de aprendizaje en cierta materia ($\bar{x}$) que también aparece al azar.

Más al respecto, en el caso de una calificación de alguna materia puede ir desde un 10% a un 90% de aciertos, otras cosas siendo iguales (i.e., céteris páribus es una locución latina que significa que las demás variables permanecen constantes; es decir, no intervienen). En este rango, las calificaciones serian representadas en la línea horizontal donde cada una de estas calificaciones tendrían cierta probabilidad representada por cierta área bajo la curva normal (Figura 3.1). La Figura 3.1 muestra dos áreas en color gris. La primera va de una calificación de 10% a 20%, y la segunda va de 50% a 60%. Como el área bajo la curva es mayor de 50% a 60% que la de 10% a 20%, es más probable que una calificación se encuentre en este rango de 50% a 60% que en el de 10 a 20%. A esto también se le llama tener más densidad: del 50% al 60%.

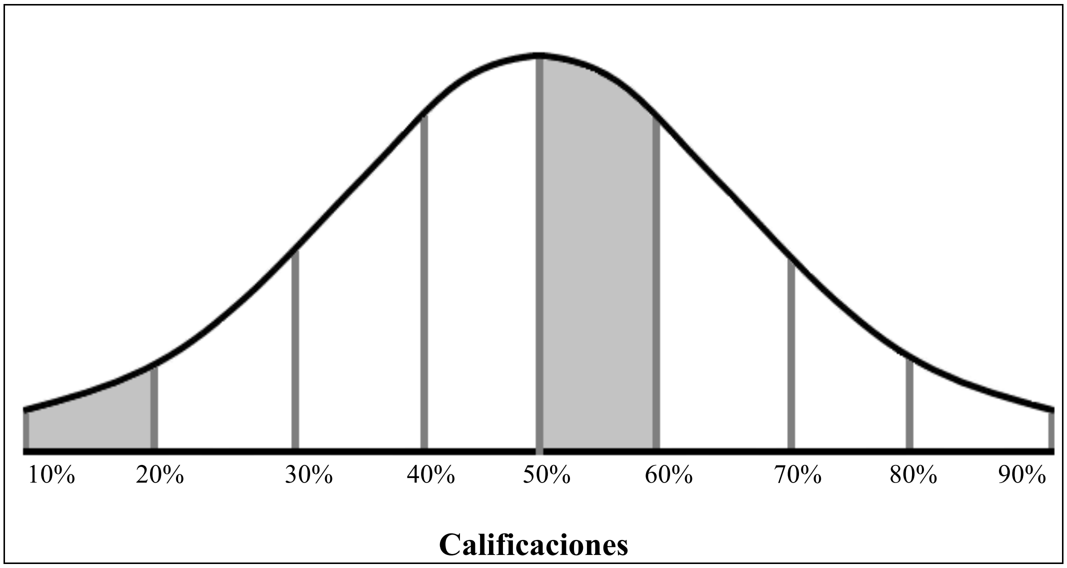
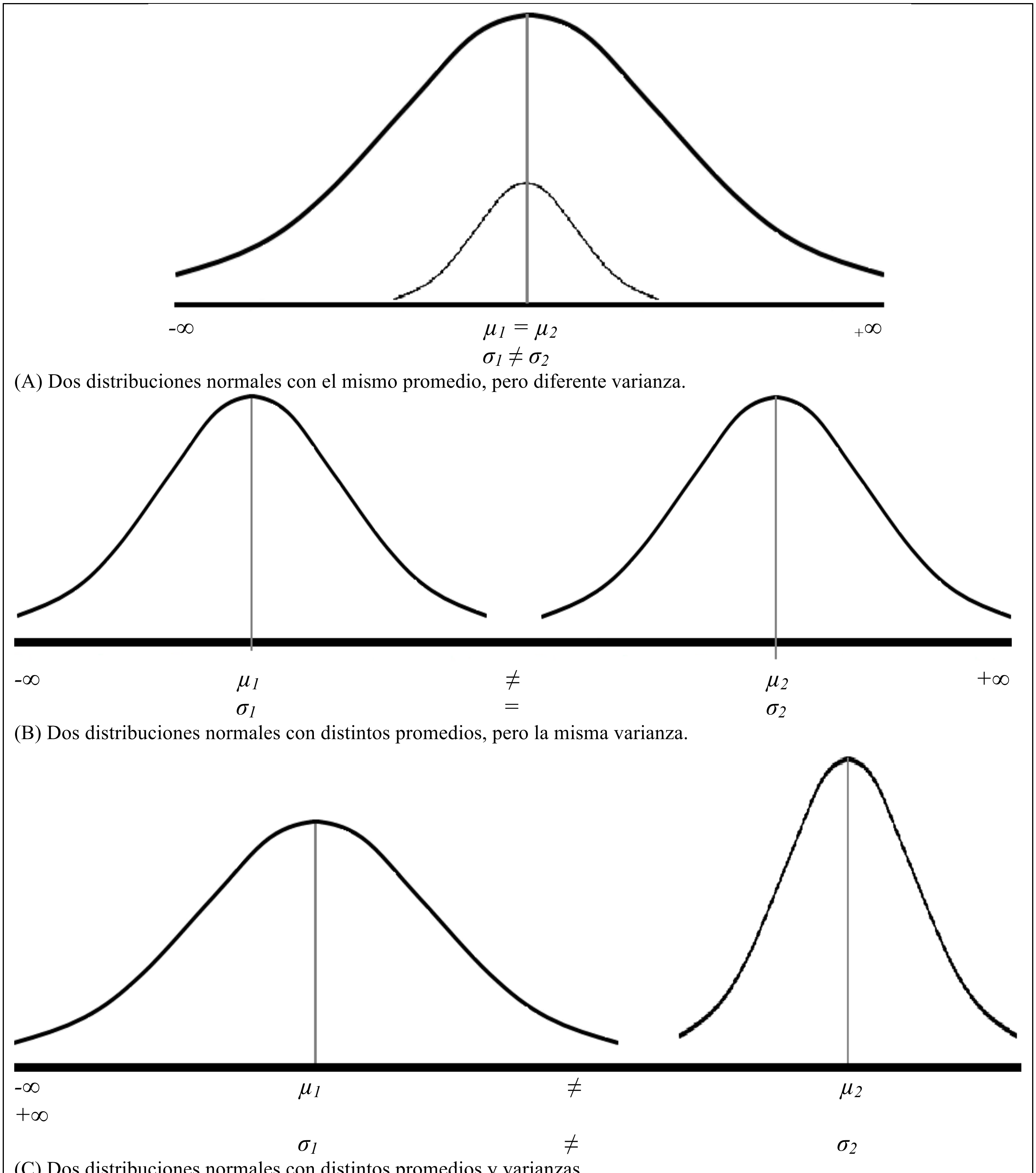


Figura 3.1 **Calificaciones en la Curva Normal.**

La Figura 3.2 muestra como la distribución normal depende del promedio y de la varianza: N (μ, σ^2). Al tener diferencias entre los promedios y varianzas de los sets de datos se tiene una familia de distribuciones normales. La Figura 3.2 muestra tres posibilidades cuando se comparan solamente dos sets de datos a la vez. Se podrían comparar un sin fin de set de datos. En el caso (A), se aprecia en los dos sets de datos que coinciden en tener el mismo promedio, pero, por otro lado, la varianza es distinta. Esto último ocasiona que una de las distribuciones sea más amplia que la otra. Para el caso (B), los promedios son diferentes, pero las varianzas son iguales. El resultado de lo anterior es que las dos distribuciones tienen la misma forma, pero hay una distancia entre ambas ya que sus promedios no están en el mismo punto. El último caso presentado en esta tabla fue el (C) que muestra que tanto el promedio como la varianza son diferentes entre las dos curvas normales. Visualmente, los promedios no coinciden en la horizontal y la varianza hace que una de las distribuciones sea más estrecha que la otra.



(A) Dos distribuciones normales con el mismo promedio, pero diferente varianza.

(B) Dos distribuciones normales con distintos promedios, pero la misma varianza.

(C) Dos distribuciones normales con distintos promedios y varianzas.

Nota: N (μ, σ^2): μ = promedio; σ^2 = varianza.

Figura 3.2 **Distribuciones Normales con sus Promedios y Varianzas.**

b. Teorema de Tendencia Central.

Uno de los pilares de las estadísticas inferenciales es el *Teorema de Tendencia Central* (*Central Limit Theorem*). Una definición de este teorema fue dada por Hinkle et al. (2003, pp. 733-734):

> Un teorema que provee una base matemática para usar la distribución normal como la distribución de muestras cuando se tiene cierto tamaño de la muestra. El teorema explica que esta distribución de muestras esta normalmente distribuida, tiene un promedio igual al promedio de la población (μ) y tiene una varianza igual a la varianza entre el tamaño de la muestra: σ^2 / n.

Para complementar esta definición, Schumacker y Tomek (2013, p. VIII) explicaron que la *distribución de muestras* (*sampling distribution*) estaría normalmente distribuida sin importar la forma de la distribución de la población, de la cual se obtuvo una muestra o más muestras. Más al respecto, Tabachnick y Fidell (2012) explicaron que el teorema dice que si se *aumenta* el tamaño de las muestras la distribución de muestras de una estadística (e.g., promedio) se distribuiría normalmente, aunque la distribución de la población no lo fuera. En otras palabras, la *distribución de muestras* de un promedio de una población que no es normalmente distribuida será aproximadamente normal, pero una muestra más grande podría ser necesitada dependiendo de la anormalidad de la población (Schumacker y Tomek, 2013, p. 96).

Por ejemplo, *equis* (x) es una variable al azar como en el ejemplo anterior fueron las calificaciones con su distribución de N (μ, σ^2). Ahora esta variable x (calificaciones) tiene un promedio (e.g., el 50%; Figura 3.1.) y viene de una población (N). Si se tomara un *número grande* de muestras de calificaciones de un *tamaño grande* (i.e., ni el número ni el tamaño de las muestras ha sido especificado en la literatura que se revisó para el presente libro), se obtendría una *distribución de muestras* como la Figura 3.3. Es decir, una distribución de muestras del promedio en forma normal o aproximada a esta. Lo anterior se puede representar (Ecuación 3.1):

$$\mathrm{N}_{\text{muestras}} (\mu, \sigma^2 / N) \qquad \text{Ecuación (3.1)}$$

$\mathrm{N}_{\text{muestras}}$ = distribución de muestras del promedio.

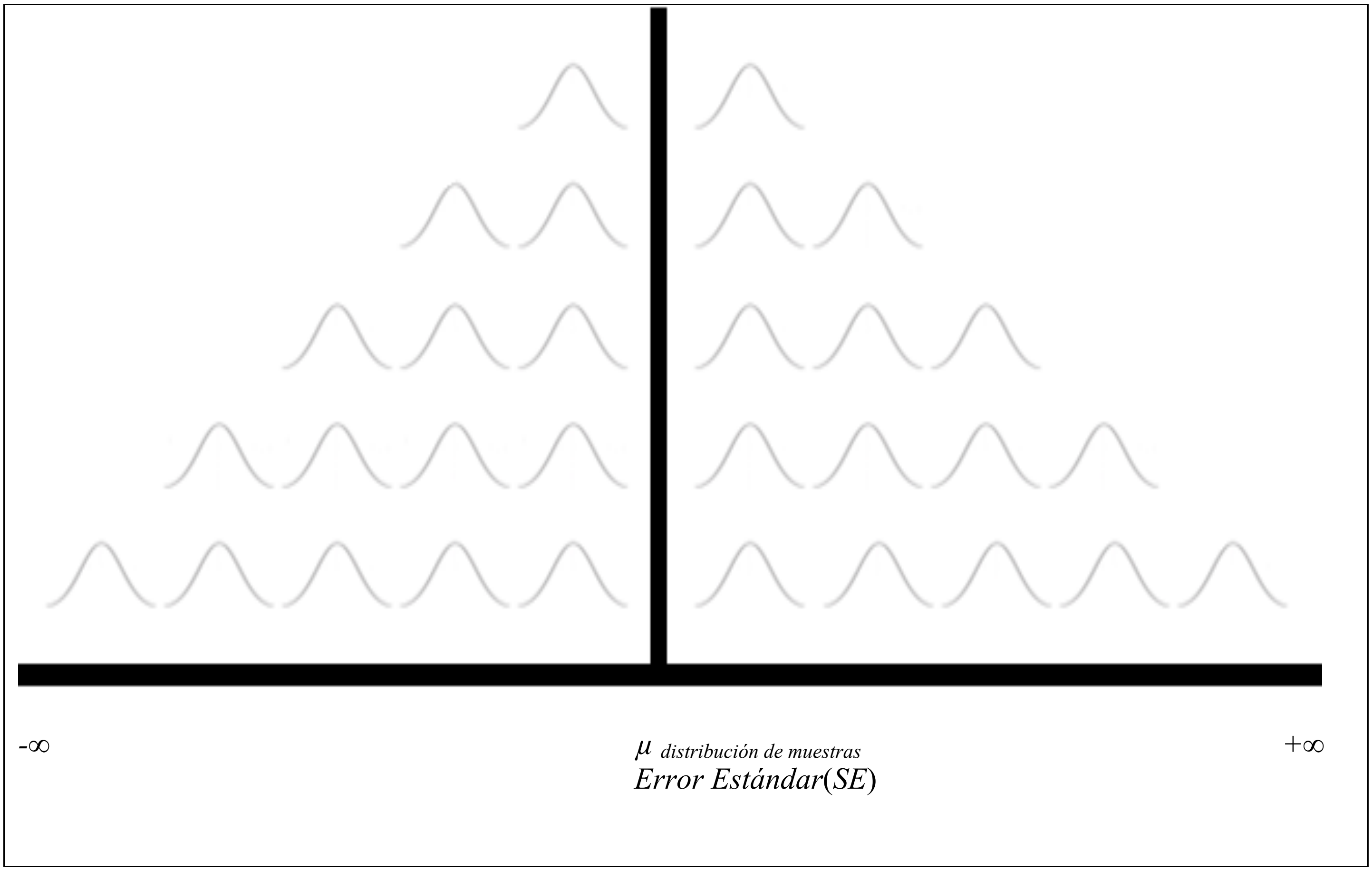
μ = promedio de la población.

σ^2 = varianza de la población.

N = Tamaño de la población

Nota: Lo anterior significa que la distribución de muestras del promedio depende del promedio (μ), la varianza (σ^2) y el tamaño de la población (N).

Cumming (2013, p. 59) explico que "el Teorema de Tendencia Central y la distribución normal sí parecen expresar algunos aspectos básicos de cómo el mundo natural funciona." Este autor dio varios ejemplos para apoyar la relación entre el mundo natural y el Teorema de Tendencia Central. En uno de ellos, él describió que las hormigas adultas de cierta especie tienen cierta longitud. Esta longitud puede ser afectada por un sin fin de variables (e.g., dieta, clima, orografía, etc.). Sin embargo, si se toma una gran cantidad de muestras de gran tamaño, la distribución de la longitud de las hormigas tendera a ser normal. Con esta distribución normal de las muestras se podrá estimar el promedio de la población.



Nota: Las equis representan las muestras de la distribución de muestras. La desviación estándar de la distribución de muestras es el Error Estándar (*SE*), el cual se explica más adelante con detalles. Asimismo, la $\mu_{\text{distribución de muestras}}$ = se promedian las muestras: i.e., promedio de la distribución de muestras. Además, el *Error Estándar* (*SE*) = Es un análogo de la desviación estándar para la distribución de muestras. Esto es: Para las muestras: $SE = SD / \sqrt{n}$ y Para las poblaciones: $SE = \sigma / \sqrt{n}$.

Figura 3.3 **Distribución de Muestras.**

Al tomar varias muestras de una población, la *distribución de esta* forma una campana de Gauss, y esta es la distribución de muestras (Figura 3.3). Cada una de las equis dentro de la Figura 3.3 representa una muestra. Por simplicidad, solo se utilizó un número limitado de muestras para ilustrar el ejemplo. La distribución de muestras es normal a pesar de que provenga de una población que tuviera otro tipo de distribución (e.g., uniforme, exponencial, bimodal o desconocida; ver Figura 3.4).

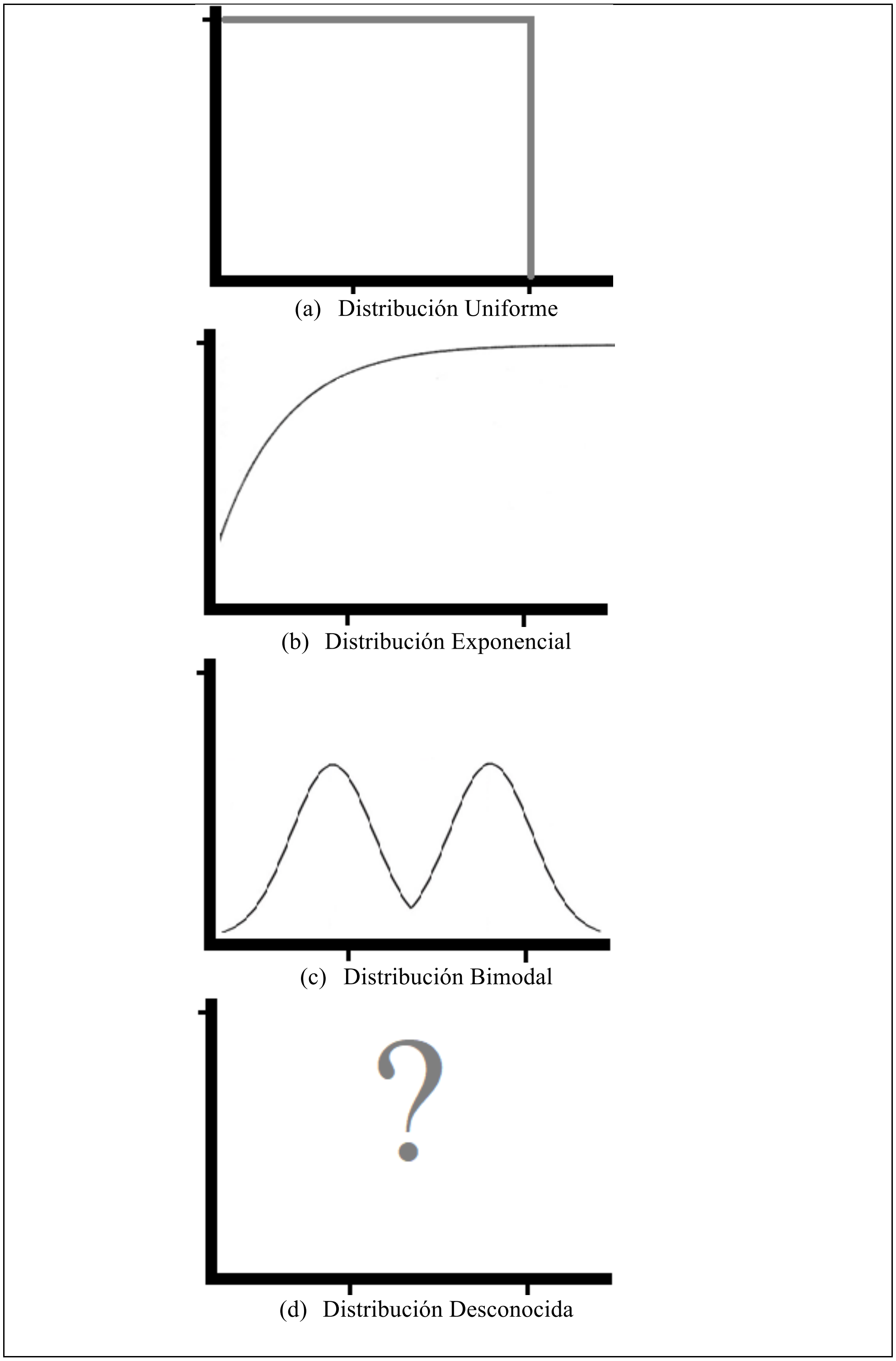


Figura 3.4 **Otras Distribuciones.**

Por ejemplo, una investigadora puede preguntarse: ¿Cuál es la distribución del promedio de los puntajes promedio de una población de interés?; y ¿Cuál es la forma de la distribución de muestras de esa población? La distribución de esta población podría tomar diferentes formas (e.g., uniforme, exponencial, binomial, o simplemente desconocida). Estas distribuciones aparecen en la Figura 3.4. La respuesta a ambas preguntas es que no importa que forma tenga la distribución de la población porque el Teorema de Tendencia Central muestra que, si la investigadora obtiene cierto número de muestras de gran tamaño, la distribución de estas muestras será normal. En otras palabras, la mayoría de las muestras tendera a acercarse al promedio de la distribución como en la Figura 3.3. Esta distribución tendrá una desviación estándar que se llama: error estándar (*standard error*; i.e., su abreviación es *SE*). La fórmula es (Ecuación 3.2):

$$SE = \frac{SD}{\sqrt{n}} \qquad \text{Ecuación (3.2)}$$

Donde *SD* = desviación estándar de una muestra.

n = tamaño de una muestra.

Dada esta fórmula, el tamaño de la muestra y el error estándar tienen una relación inversa: cuando uno incrementa el otro disminuye. Lo anterior implica que menos error estándar es asociado con un tamaño de la muestra mayor. Un ejemplo sería para un estudio de puntajes de inteligencia porque en este tipo de test se tiene una desviación estándar de 15 puntos. La Tabla 3.1 muestra esta desviación estándar de un test de inteligencia de 15 puntos que se mantiene fija cuando la muestra va aumentando de 10 a 100 y a 1,000. Realizando las operaciones, se obtiene los diferentes errores estándar que van disminuyendo conforme la muestra aumenta (la flecha de la Tabla 3.1. muestra el sentido en que disminuye los coeficientes de *SE*).

Tabla 3.1 Relación Inversa del Errores Estándar y el Tamaño de la Muestra.

SD	*N*	$\sqrt{n}$	$SE = SD / \sqrt{n}$
15	10	3.16	4.75
15	100	10	1.5
15	1,000	31.62	0.47

Una recomendación para los investigadores es que traten de obtener la muestra más grande que se pueda porque esto ayuda a calcular intervalos de confianza más precisos al disminuir el *SE*. Habría que también considerar los recursos que se tienen para llevar a cabo un estudio porque estos pueden limitar que tan grande seria la muestra. En la siguiente sección se cubren los detalles de los intervalos de confianza.

De nuevo, al usar el Teorema de Tendencia Central se forma una distribución de muestras. Cuando se usa la distribución de muestras para llevar a cabo comparaciones entre grupos como en un test *t* de estudiante se usa la distribución de muestras de las diferencias. Esta distribución de muestras de las diferencias es análoga a la distribución de muestras. Sin embargo, ambas distribuciones tienen la misma forma, promedio y varianza, así que se puede usar una Tabla *t* para calcular valores críticos de las dos porque serán los mismos. Para los propósitos del presente libro, no se distinguirá entre la distribución de muestras y la distribución de muestras de las diferencias. Para una discusión más a fondo de la distribución de muestras de las diferencias, se recomienda a Hinkle et al. (2003).

c. Estadísticas Inferenciales para las Muestras.

Para entender mejor las estadísticas, hay una distinción entre estadísticas y parámetros. Las estadísticas son estimaciones de algún atributo/propiedad de una población mientras los parámetros son los atributos/propiedad de una población (Schumacker y Tomek, 2013). Las estadísticas descriptivas no involucran la capacidad de generalizar las características al resto de la población (Lane et al., 2014).

Uno podría ver a las estadísticas inferenciales, en general, como una colección de métodos para hacer inferencias de una muestra a la población (Kotz, 2006, p. 3411). Como no se tiene a toda la población con una o varias variables de interés, Kotz (2006) considero a una muestra como si fueran datos incompletos. De hecho, suele suceder de que un investigador o investigadora educativa encuentre que le faltan datos por alguna razón. Para remediar esta situación de valores faltantes se recomienda a Enders (2010) para usar uno de sus métodos de imputación de datos. Hinkle et al. (2003, p. 738) definieron a una población y una muestra:

> Una población: Todos los miembros de un grupo definido (e.g., estudiantes de universidad en Europa). Puede estar formada de personas, observaciones u objetos.
>
> Una muestra: Un subset de una población (e.g., una porción de la población de estudiantes de universidad en Europa). Asimismo, puede estar formada de personas, observaciones u objetos.

Volviendo al tema, una muestra (i.e., subset) es usada para hacer inferencias sobre la población (i.e., set). Más en detalle, se obtiene una muestra que es una parte de una población, se mide una o varias estadísticas (e.g., el promedio), se estima un error de muestreo (i.e., diferencia entre el valor de una estadística de una muestra y un parámetro de la población), y se hace una inferencia que cierto parámetro tiene tal valor, céteris páribus. De nuevo, el error de muestreo para el promedio se representa aquí:

Error de muestreo del promedio = μ - $\bar{x}$

Donde μ es el promedio de una población.

$\bar{x}$ es el promedio de una muestra.

Si se tiene la información de toda la población, se llama *censo*, y no es necesario usar las estadísticas inferenciales. En un censo no existe un error de muestreo y se medirían directamente los parámetros de una población (i.e., μ, σ^2), sin necesidad de hacer inferencias desde una muestra. Los parámetros de una población (μ, σ^2) son valores fijos, pero desconocidos (Cumming, 2013, p. 55).

Las estadísticas inferenciales se basan en el supuesto que se *muestreo al azar* una población, en la cual cada uno de sus integrantes tenía la misma oportunidad de ser seleccionada o seleccionado (ver el *Muestreo Simple al Azar* Tabla 3.2). También, la selección de cada miembro debe de ser independiente de la selección de otro miembro. Esto quiere decir que la selección de algún miembro de la población no debe de disminuir o incrementar la probabilidad de elegir a otro miembro. El tamaño de la muestra importa. Entre más grande la muestra es mejor porque se aproximaría al tamaño de la población. Por ello, Lane et al. (2014) dijeron que solamente muestras grandes hacen probable su representatividad.

Existen otros tipos de muestreos más complejos y Berenson, Levine y Krehbiel (2012) presentaron una serie de métodos para muestrear. Un ejemplo de esto es un método que puede usar cuando la población esta estratificada (i.e., con subgrupos). Entonces, se trata de tomar muestras de estos subgrupos que serán proporcionales al tamaño de los grupos con respecto a la población. Ejemplo, si la población tiene un 70% de varones y 30% de damas, la muestra también tendrá este 70% de varones y 30% de damas (ver Muestreo Estratificado al Azar en la Tabla 3.2.).

d. Muestreo.

Las muestras se suelen usar porque los investigadores educativos no cuentan con los recursos para llevar a cabo un censo. Esto no solo afecta a los investigadores educativos (i.e., presupuesto relativamente limitado) sino también al resto de los investigadores. Por ello, se opta por una o varias muestras que representen determinada población. El tema del muestreo puede llegar a ser complejo y muchos libros han sido dedicados exclusivamente a este tema. Un libro de Sudman (1976; i.e., *Applied Sampling*) sobre el muestreo podría ampliar lo que se expone en esta sección. Asimismo, la Tabla 3.2 muestra algunas formas de muestrear. El tamaño de la muestra viene después de las formas.

Tabla 3.2 Tipos de Muestreo.

Tipo de Muestreo	Implicaciones
Muestreo Simple al Azar	Se debe de tener acceso a cada miembro de la población de interés. Con ello, cada miembro de esta población tiene la misma posibilidad de ser seleccionada o seleccionado. La selección de una o un miembro es independiente de la selección de otra u otro. Al seleccionar un miembro esto no disminuye o aumenta las probabilidades de seleccionar a alguien más.
Muestreo sin ser al Azar	Sucede cuando se tiene una muestra de conveniencia o cuando se seleccionó al azar una muestra, pero los miembros de esta muestra optaron por no participar. Entonces, solo quedaron aquellos y aquellas que por algún motivo se autoseleccionaron para participar en un estudio. Suele pasar en la investigación educativa que en los estudios participen solo aquellos y aquellas que tienen algún interés, y un investigador o investigadora debería de tener cuidado con las conclusiones que sacaría de resultados de estos voluntarios.
Muestreo Sistemático	Pasa cuando se tiene una lista de los miembros de la población, y mediante algún método se selecciona un número que servirá para elegir a los participantes. Un ejemplo sería una lista de 6,000 personas (*N*). Se tienen los recursos para sacar una muestra de 100 (*n*) participantes. Se hace la siguiente operación: n / N = 100 / 6,000 = 1 / 60 Una de cada 60 personas seria elegida. Ahora habría que elegir un número entre 1 y 60 al azar para que sea el punto de inicio. Suponiendo que se eligió el número 13, la primera persona seleccionada de la lista sería: la 13ª. Entonces, la segunda seria 13 + 60 = 73ª; la cuarta… 73 + 60 = 133ª. De este modo, hasta llegar a 100 participantes.
Muestreo de *Clúster*	Este tipo de muestreo involucra también el tener una lista de la población para hacer una selección al azar. La diferencia con las anteriores es que los participantes se encuentran anidados en grupos/clústeres, de los cuales resultaría difícil separarlos, pero los grupos/clústeres si pueden ser seleccionados al azar. Un ejemplo de investigación educativa seria que se desea tener una muestra al azar para un estudio de corte experimental. Este estudio implica el uso de un tratamiento dentro de las horas de clase. Es decir, tendría que tomar lugar en un salón de clases sin que se pudieran separar a los estudiantes de este salón. Entonces, los salones de clases serían los clústeres que se seleccionarían al azar.
Muestreo Estratificado al Azar	Aparece cuando la población se encuentra estratificada. Es decir, se sabe que subgrupos la integran (e.g., sexo, edad, grupo social, nivel socioeconómico, etc.). Dado este conocimiento, se puede seleccionar en proporción a estos estratos/grupos. Por ejemplo, si la población tiene un 70% de damas y un 30% de caballeros, la muestra mantendrá estos porcentajes también.

Nota: cf. Hinkle et al. (2003). Una calculadora útil para estimar el tamaño de muestras es *Sample Size Calculator* de *Creative Research Systems*: https://www.surveysystem.com/sscalc.htm

Hay diferentes maneras de estimar el tamaño de muestra. Estas maneras dependen de la información que se tenga en el momento de calcular la muestra. Otro aspecto para considerar en lo práctico es los recursos con los que se cuente para llevar a cabo el estudio. La Tabla 3.3 muestra algunos aspectos a considerar a la hora de estimar el tamaño de la muestra:

Tabla 3.3 Variables Asociadas al Tamaño de la Muestra.

Variable	Definición
Tamaño de la Población	Habría que partir de una definición de la población que se tiene en mente estudiar: objetos, observaciones o personas (e.g., edificios, calificaciones o estudiantes de escuela elemental, respectivamente). Esto también con el propósito de poder generalizar los resultados de la muestra a la población. En el mejor de los casos, se suele hacer una aproximación al tamaño de la población. En otras ocasiones, el tamaño de la población es desconocida. Para tener una idea de este tamaño, habría que consultar censos, estimaciones nacionales e internacionales.
Margen de Error/ Intervalo de Confianza	Se dan más detalles en la próxima sección acerca de estos conceptos de margen de error e intervalo de confianza. El supuesto es que hay una diferencia (i.e., error) entre un parámetro de una población (e.g., promedio) y la estadística de su muestra correspondiente. Dado este supuesto, se calcula cuanto error es aceptable en porcentaje y su correspondiente valor z: e.g., 1% ($z = 2.58$), 5% ($z = 1.96$), o 10% ($z = 1.65$). El error aceptable depende la precisión que se deseen de los resultados: más precisión implica menos error. Tradicionalmente, un error del 5% es aceptable en la investigación educativa, pero habría que consultar la literatura del tema en específico para tener una idea informada del tamaño aceptable del error.
Varianza/ Desviación estándar	Esta varianza puede ser obtenida de estudios previos en forma de *SD*. Si no los hay, algunos investigadores han recomendado el 0.5 para usarla como *SD* en las fórmulas de estimación del tamaño de una muestra.
El análisis que se va a utilizar.	También, habría que considerar los análisis estadísticos y psicométricos a utilizar para analizar los datos. Por ejemplo, si se van a utilizar análisis estadísticos como comparaciones entre grupos o relaciones entre variables, una opción es consultar a Tabachnick y Fidell (2012). También, el análisis psicométrico implica saber los requisitos del tamaño de la muestra de una teoría psicométrica: Teoría Clásica del Escore Verdadero (*Classical True Score Theory*; ver a Crocker y Algina, 2006); Teoría de la Respuesta al Ítem (*Item Response Theory;* Embretson y Reise, 2000); y la Teoría de la Generalización (*Generalizability Theory*; Shavelson y Webb, 1991)
Poder estadístico	Este tema se tratará más adelante con más detalles. Por el momento, este poder estadístico es una probabilidad que depende en parte del tamaño de la muestra. Para la planeación de un estudio, el poder estadístico sirve para calcular el tamaño de una muestra y tener el suficiente poder para rechazar una hipótesis nula cuando es falsa (cf. Cohen, 1988; Salkind, 2007). Es decir, se desea el poder rechazar una hipótesis nula cuando sea falsa. Una hipótesis nula es falsa cuando se tomaron un par de muestras para compararlas y al usar un análisis de comparación de promedios (e.g., *t* test o análisis de la varianza) resultó que no había una diferencia entre ambos promedios. Sin embargo, cuando se compararon los promedios de las poblaciones de los cuales venían estas muestras, se observó que si había una diferencia entre ambos. En corto, no se rechazó la hipótesis nula con los promedios de las muestras cuando la hipótesis nula era falsa y esto se apreció hasta tener los promedios de las poblaciones de origen.

Existen muchas fórmulas y recomendaciones para la estimación de una muestra. Estas dependen también en que estadística (e.g., promedio, *SD*) se quiere usar tanto como el tipo de análisis estadístico que se use (e.g., *t* test, análisis de la varianza, etc.). También, el implementar procesos psicométricos de confiabilidad y validación implican el considerar el tamaño de la muestra (Ponce, aceptado y en prensa). Asimismo, existen varias calculadoras en páginas de estadísticas para estimar el tamaño de la muestra (e.g., la mencionada en la Tabla 3.2). Una fórmula para estimar el tamaño de la muestra cuando se quiere generalizar el promedio de una muestra a la población es la siguiente (Ecuación 3.3):

$$\text{Tamaño de la muestra } (n) = \frac{z^2 \text{ X } SD \text{ X } (1 - SD)}{(\text{Margen de error})^2} \qquad \text{Ecuación (3.3)}$$

z = Este valor z especifica cuando error es aceptable para el estudio (2.58, 1.96 y 1.65, entre muchos otros).

SD = La desviación estándar de algún estudio previo o, en su defecto, 0.5.

Margen de error: Expresado en porcentaje que corresponda a los valores z de 2.58, 1.96 y 1.65, respectivamente (i.e., 1%, 5% y 10%). Para el uso de la formula estos porcentajes se convierten en decimales: i.e., .01, .05 y .10.

Nota: En esta estimación el tamaño de la población es desconocida. Hay otras fórmulas que consideran el tamaño de la población para estimar las muestras.

Usando la Ecuación 3.3, un ejemplo heurístico con una desviación estándar de 0.5, intervalo de confianza del 99% y margen de error (z = 2.58 y 1%) es:

$n = [(2.58)^2 \text{ X } 0.5 \text{ X } (1 - 0.5)] / (.01)^2$

$n = [6.66 \text{ X } 0.25] / .0001$

$n = 1.665 / .0001 = 16{,}650$

Para obtener un margen de error del 1%, se necesita una muestra de un tamaño aproximado de 16,650 (e.g., participantes, observaciones u objetos). El Capítulo 4 del presente libro explica más a fondo el tema de los intervalos de confianza.

Preguntas para Reflexionar

1.- ¿Cuáles son las dos variables más importantes para la distribución normal de un set de datos?

2.- ¿Por qué lo valores más cercanos al promedio son más probables que el resto?

3.- ¿Cómo se podría explicar el Teorema de Tendencia Central a personas que no están muy familiarizas con conceptos estadísticos?

4.- ¿Cómo se podría utilizar el Teorema de Tendencia Central para defender una muestra que proviene de una población de la cual se sospecha que no tiene una distribución normal?

5.- ¿Cuáles podrían ser las razones de más peso para usar cierto tipo de muestreo?

6.- ¿Cuáles son los riesgos de usar muestras donde no se usó el azar para seleccionar los participantes, observaciones u objetos?

7.- ¿Qué variables afectan el tamaño de una muestra?

Opinión del Autor

Lo esencial de este capítulo es la comprensión del Teorema de Tendencia Central que explica que cuando se toma un gran número muestras de tamaño grande se obtiene una distribución de muestras que es similar a una distribución normal. Desafortunadamente, no se especificó ni el número de muestras ni el número específico para lograr esta distribución de muestras similar a la distribución normal. Pese a esta incertidumbre y dado que los y las investigadoras tienen recursos limitados, se pueden usar otras variables para precisar el tamaño de la muestra. Entre estas variables están: el tamaño de la muestra; margen de error; varianza de los datos; el análisis para ser usado; y el poder estadístico.

Capítulo 4. Intervalos de Confianza.

> Posiblemente, uno ha escuchado términos como el margen de error cuando llegan los tiempos de elecciones políticas. Con encuestas a los posibles votantes, se trata de anticipar el resultado de unas futuras elecciones. En otras palabras, se trata de hacer una predicción con los resultados de estas encuestas. Lo complicado de hacer una predicción es que las personas encuestadas puede que no sean sinceras o que cambien de opinión una vez que lleguen a las urnas a depositar su sufragio. Si se asume una curva normal sobre las preferencias políticas, se tendrían dos colas. Una cola que siempre votaría por un partido o una o un candidato, y, por otro lado, la otra cola de la curva nunca votaría por el partido o una o un candidato. Entre las colas está el resto de los votantes que pueden elegir a un partido u al otro, y, lo mismo, sucedería por los candidatos o candidatas. Para hacer inferencias de quién ganará una elección se usan los intervalos de confianza. Con estos intervalos, se puede estimar el promedio de una población. También, se pueden estimar muchos otros parámetros de una población como su varianza, coeficientes de confiabilidad, etc.

Según Salkind (2007, p. 177), la estimación de intervalos de confianza (*confidence interval*; *IC*) con cierto nivel de confianza lleva a capturar el verdadero parámetro de la población en cierto porcentaje de las muestras. Es decir, si se quiere un nivel de confianza del 99% para los intervalos de confianza de un promedio de población (μ), esto implica que de cada 100 intervalos de confianza que se calculen, 99 de ellos capturan el promedio verdadero de la población y uno no lo hará. Al respecto, Schumacker y Tomek (2013, p. 142) dijeron que se puede calcular un *IC* para muchos de los parámetros de una población (e.g., promedio, desviación estándar, etc.) al usar el error estándar de la estadística de interés y el nivel de confianza. También se pueden calcular un *IC* para coeficientes de correlación (Producto-momento de Pearson) y de confiabilidad (e.g., Alfa de Cronbach). Para los temas de *IC* para coeficientes de correlación y confiabilidad se recomienda a Thompson (2008) y a Thompson (2003) respectivamente. Los intervalos de confianza más comunes son del

68%, 95% y 99%. Algunos autores recomiendan un intervalo del 95% para la investigación educativa. Para calcular el intervalo de confianza del promedio de la población se usa la siguiente Ecuación (4.1):

Un intervalo de algún nivel de confianza: $\bar{x}$ +/- z_{critico} X $\frac{SD}{\sqrt{n}}$ Ecuación (4.1)

$\bar{x}$ = Promedio de una muestra.

z_{critico} = Nivel de confianza que se desee (68%, 95%, o 99%) expresado con un valor *z* (i.e., 1.65, 1.96, 2.58, respectivamente).

X = Multiplicación

SD = Desviación estándar de la muestra.

n = Tamaño de la muestra.

Nota: Recordando, el error estándar ($SE = SD / \sqrt{n}$)

Como indica la parte de la formula con los signos de más y de menos (+/-), el intervalo tendrá en el centro al promedio y será simétrico porque se le suma y se le resta la misma cantidad. La Tabla 4.1 muestra como calcular un set de intervalos de confianza de los más comunes (68%, 95% y 99%) para un test de inteligencia con un $\bar{x}$ = 100, una *SD* = 15, y un *n* = 150. Se obtiene el nivel de confianza de la Tabla de los valores *z* del Apéndice.

Tabla 4.1 Tabla del Cálculo de Intervalos de Confianza.

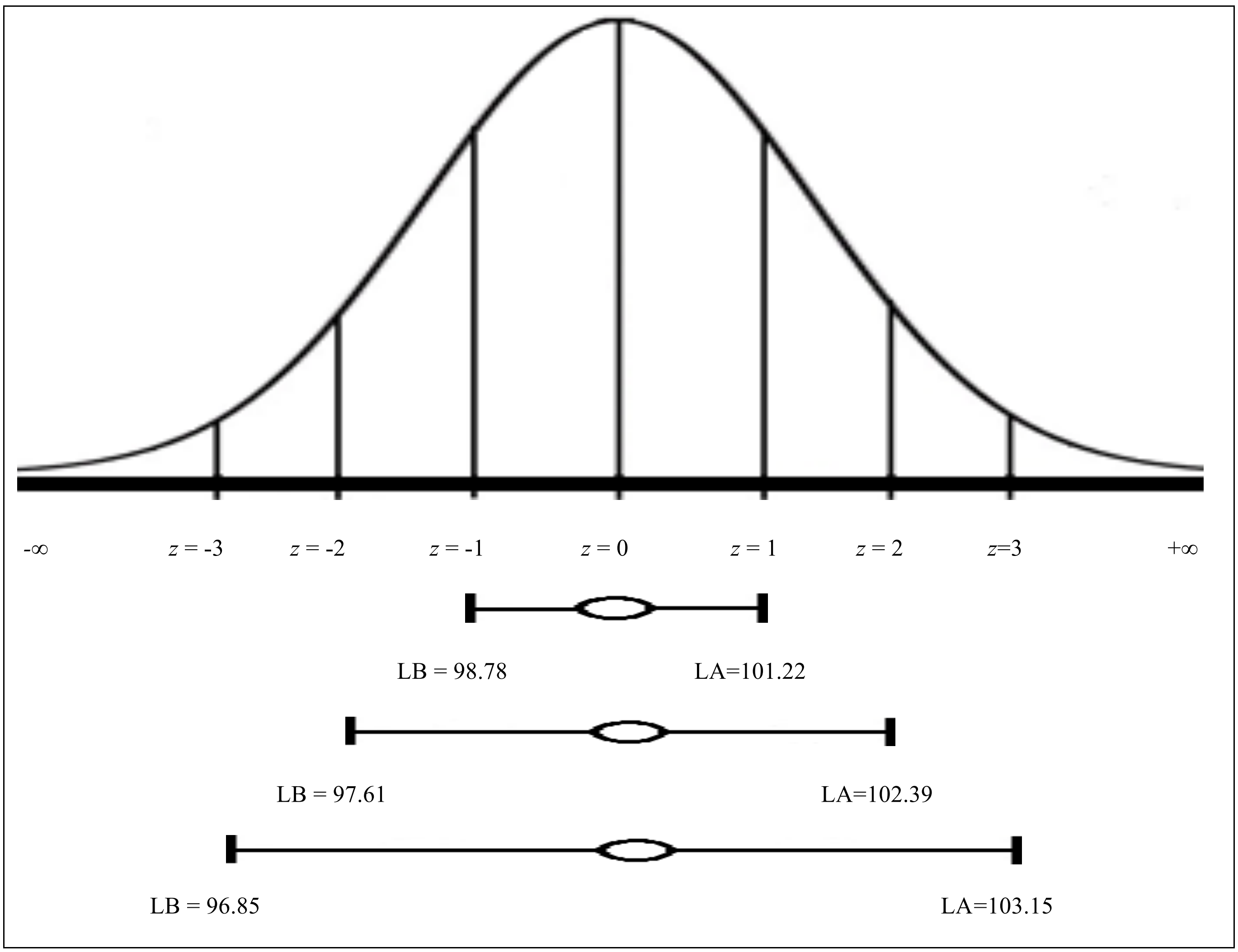
Nivel de Confianza	Valor Critico *z*	*SD*	*n*	z_{critico} x $SD / \sqrt{n} = a$	$\bar{x}$ - a (Limite bajo)	$\bar{x}$ + a (Limite alto)
68%	1*	15	150	1 x 15 / $\sqrt{150}$ = 1.22	100 – 1.22 = 98.78	100 + 1.22 = 101.22
95%	1.96**	15	150	1.96 x 15 / $\sqrt{150}$ = 2.39	100 – 2.39 = 97.61	100 + 2.39 = 102.39
99%	2.58***	15	150	2.58 x 15 / $\sqrt{150}$ = 3.15	100 – 3.15 = 96.85	100 + 3.15 = 103.15

Nota: Para las probabilidades asociadas con los valores críticos: * en la tabla *z* tiene un valor aproximado a .3413; ** valor aproximado a .4750; y *** valor aproximado a .4951. Por motivo del espacio de la Tabla, *a* representa la ecuación: z_{critico} x $SD / \sqrt{n}$ con la letra *a*. El error estándar (*SE*) es de 1.22 para los tres niveles de confianza porque la *SD* y el tamaño de la muestra son constantes.

La Figura 4.1 muestra la representación gráfica de estos intervalos de confianza de la Tabla 4.1: (i.e., 68%, 95%, y 99%). Se puede apreciar que los intervalos de confianza con un nivel de confianza mayor tienen más longitud más amplia (i.e., el de *IC* de 99% > *IC* de 95% > *IC* de 68%). La longitud que esta del promedio hacia la derecha, se llama limite alto (*upper limit*) o cola superior del *IC*. La otra parte de la longitud que esta del promedio hacia la izquierda es el límite bajo (*lower limit*) o cola inferior del *IC*. Estos límites se calculan de la siguiente manera de la Ecuación (4.1):

Limite alto (LA) = $\bar{x}$ + ($z_{\ critico}$ X SD / $\sqrt{n}$)

Limite bajo (LB) = $\bar{x}$ - ($z_{\ critico}$ X SD / $\sqrt{n}$)



Nota: Las abreviaciones LB = Limite bajo del intervalo de confianza; y LA = Limite alto.

Figura 4.1 **Intervalos de Confianza de la Tabla 4.1.**

La Figura 4.2 muestra un ejemplo de lo antes mencionado por Salkind (2007, p. 177): captura del *valor verdadero* de la población. Este ejemplo muestra el concepto de un *IC* del 80%. En otras palabras, de un intervalo de confianza del 80% se esperaría que de 10 intervalos 8 de ellos capturarían el valor verdadero de la población y 2 de ellos no lo hicieran. No es común calcular un *IC* del 80%, pero, por la simplicidad del argumento, se calculó de este modo.

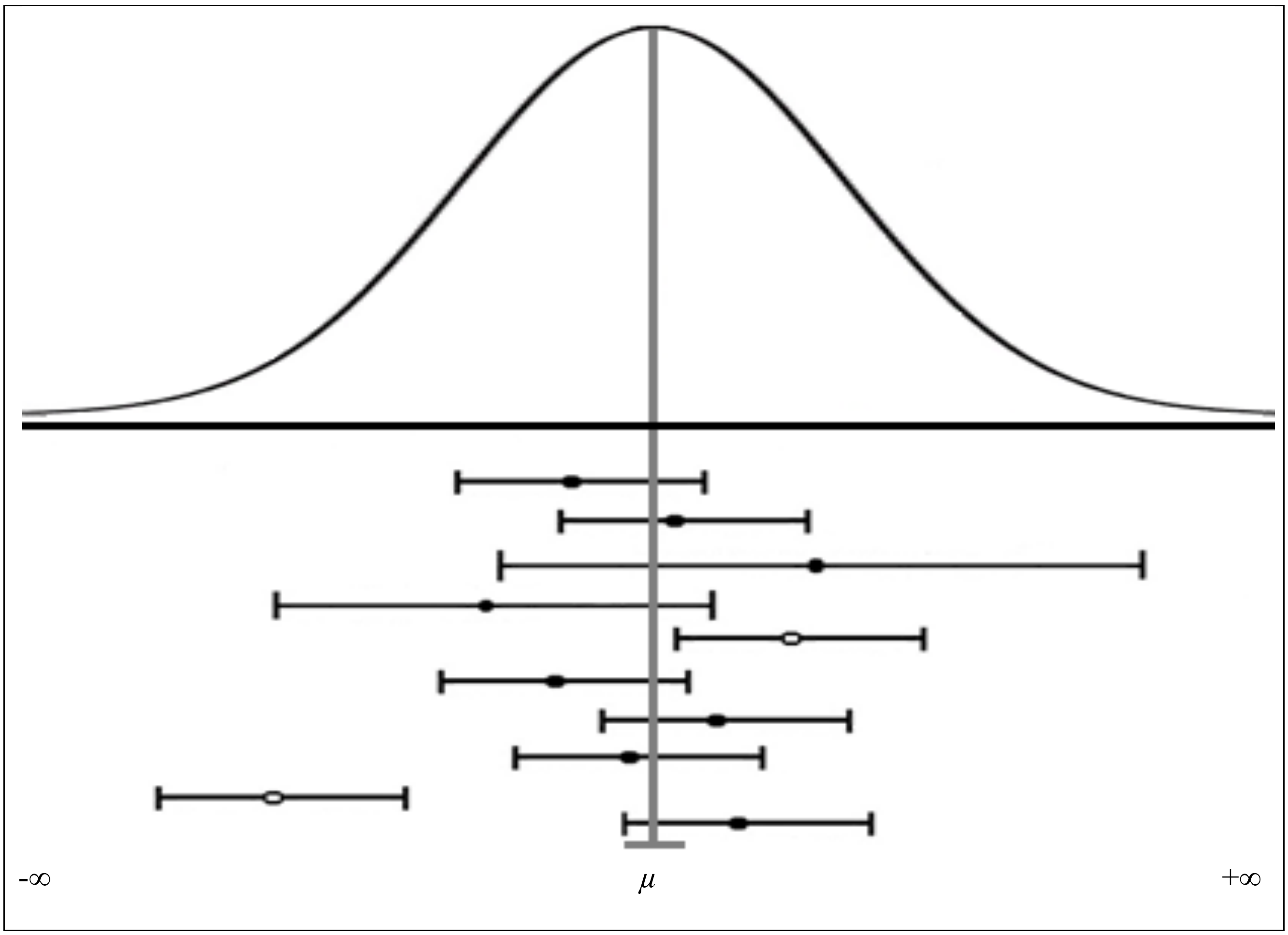


Figura 4.2 **Intervalos de Confianza del 80%.**

Los intervalos de confianza se pueden calcular cuando se trabaja con muestras y sus correspondientes estadísticas. Suele pasar que muchas y muchos investigadores simplemente reportan valores exactos de sus estadísticas como el promedio, la *SD* o el Coeficiente del Alfa de Cronbach como si estos fueran los parámetros de la población. Cuando se reportan las estadísticas sin usar intervalos de confianza, se llaman puntos de estimación (*point estimates*), y no se sabe que tan precisos son estos puntos al estimar los parámetros de una población porque no se reportaron márgenes de error. Para tener más precesión en estimar el valor real de parámetro, es necesario calcular un *IC* (cf. Cumming, 2013).

Preguntas para Reflexionar

1.- ¿Cuáles son los tipos de intervalos de confianza más comunes?

2.- ¿Qué usos se les puede dar a los intervalos de confianza en la investigación educativa?

3.- ¿Qué inferencias se pueden hacer de los intervalos de confianza?

4.- ¿Cómo se podría utilizar el Teorema de Tendencia Central para defender una muestra que proviene de una población de la cual se sospecha que no tiene una distribución normal?

5.- ¿Cuáles podrían ser las razones de más peso para usar cierto tipo de muestreo?

6.- ¿Cuáles son los riesgos de usar muestras donde no se usó el azar para seleccionar los participantes, observaciones u objetos?

7.- ¿Qué variables afectan el tamaño de una muestra?

Opinión del Autor

Espero que el lector tenga en mente que el usar muestras implica errores. Una solución para considerar estos errores es usar intervalos de confianza. Con estos, se estima tanto el error al calcular el tamaño de una muestra para que represente a una población, así como el valor verdadero de un parámetro de una población. Por lo tanto, el manuscrito de un o una investigadora estaría incompleto si solo reporta puntos de estimación sin considerar el margen de error de sus estadísticas. Existen varios autores que defienden a ultranza el uso de los *IC* en investigación educativa (Cumming, 2013), y creo que con el tiempo su uso puede generalizarse.

Capítulo 5. Estadísticas Paramétricas vs. No-paramétricas.

> En ocasiones una investigadora se encuentra con un set de datos capturados en una base de datos que ya se encuentran en cierta escala. Por ejemplo, se podría encontrar en la base que hay una variable con el nombre del estudiante y otra variable que indica si este o esta aprobó o no aprobó la clase. Si la o el investigador está planeando usar una comparación de grupos con las calificaciones como una variable continua, no se va a poder hacer porque solo va a poder obtener el número de estudiantes que aprobó o no aprobó cierta clase. En pocas palabras, solo podrían obtener las frecuencias, proporciones o porcentajes de los datos. Ante este panorama, se hubiera tenido que usar un análisis no paramétrico donde no es necesario tener variables continuas porque solo se necesitan las frecuencias. Otra situación hubiera sido, si la investigadora hubiera encontrado que las calificaciones estaban en forma de una variable continua: e.g., 8.15, 8.19, 9.01, etc. Entonces, un análisis paramétrico pudiera ser factible (entre otros requisitos para estos test) porque se estaba usando una variable continua.

Dentro de las estadísticas inferenciales se encuentran las paramétricas y las no-paramétricas. Con las estadísticas paramétricas, se obtiene de una muestra que representa a una población, la cual tiene parámetros fijos (e.g., μ y σ) y tiene cierta distribución que puede ser normal o no (cf. Schumacker y Tomek, 2013). En las estadísticas paramétricas se usa el concepto del Teorema de Tendencia Central con escalas continuas. Para las estadísticas no-paramétricas, los datos se analizan en forma de frecuencias, porcentajes y proporciones. También, se pueden utilizar datos continuos para los análisis no-paramétricos que no cumplan con los supuestos de normalidad, homogeneidad de varianzas, linealidad, tamaño de la muestra (para más información al respecto, se recomienda a Hinkle et al., 2003). De hecho, las estadísticas no-paramétricas tienen menos supuestos que las paramétricas (cf. Hinkle et al., 2003; para más información acerca de estas estadísticas se recomienda a Siegel y Castellan, 1995). La Tabla 5.1 muestra algunos de los supuestos de las estadísticas paramétricas en comparación a las no paramétricas, así como algunos ejemplos de los diferentes tipos de ambas estadísticas.

Tabla 5.1 Algunos de los Supuestos para las Estadísticas Paramétricas y No-paramétricas.

Supuestos	
Paramétrico	No-paramétrico
Unos ejemplos del uso de estas estadísticas: *t-test* (para comparar dos grupos) Correlación con el Coeficiente Producto-Momento de Pearson (para relacionar dos variables de intervalo o continuas). Regresión Lineal Simple (para predecir un valor y para explicar un fenómeno; ver a Kutner, Nachtsheim y Neter, 2004). Análisis de la Varianza (ANOVA; Maxwell y Delaney, 2004).	Unos ejemplos del uso de estas estadísticas: Chi Cuadrada (χ^2) también conocido este test como chi cuadrado para comparar frecuencias en lugar del *t*-test de grupos independientes. Wilcoxon test usado en lugar del *t*-test cuando solo se tiene una muestra. Mann-Whitney Test usado lugar del *t*-test cuando se tienen dos muestras. Kruskal-Wallis, test de Mood de la mediana usado en lugar del ANOVA de un factor.
Las muestras fueron obtenidas al azar de una población que se definió de alguna manera: e.g., la población de los estudiantes universitarios de pregrado en instituciones privados y publicas de Latinoamérica.	Se pueden usar muestras relativamente pequeñas: e.g., ver a Berenson et al. (2012) para más detalles.
El obtener una muestra es independiente de obtener otra.	Libre de limitaciones: e.g., ver a Tabachnick y Fidell (2012) para más detalles.
La variable dependiente es medida en una escala de intervalo (por lo menos). Las mediciones de intervalo implican una misma distancia entre los diferentes puntos, pero no tienen un cero absoluto: e.g., la temperatura en grados Celsius donde la distancia entre un grado y otro es la misma, pero el cero grados es arbitrario porque no implica la total carencia de temperatura. En cuestión de aprendizaje, una calificación de 7 a 8 implica la misma distancia (intervalo) de 8 a 9. Asimismo, un cero es arbitrario porque no implica necesariamente una ignorancia total.	La distribución de los puntajes típicamente se asume hipotéticamente y las distribuciones observadas son puestas a prueba en contra de las distribuciones hipotéticas.
La variable dependiente está distribuida en la población de una manera normal (homogeneidad de las varianzas): e.g., la variable dependiente es el aprendizaje y se mide operacionalmente con las calificaciones.	La variable dependiente no necesita estar distribuida normalmente en la población: e.g., la variable dependiente es el aprendizaje y se mide operacionalmente con ser clasificado como paso la metería o no la paso. Lo anterior se convertiría en frecuencias para su análisis.

Nota: Esta lista de test paramétricos y sus análogos no-paramétricos no pretende ser exhaustiva. Existen publicaciones muy completas en cuestión de estadísticas como la de Kotz (2006) mencionada a continuación que contiene muchos detalles sobre estos test.

Para ahondar en los usos y diferencias del análisis paramétrico y no paramétrico, se recomienda a la Enciclopedia de las Ciencias Estadísticas 2ª ed. (*Encyclopedia of Statistical Sciences, Second Edition*) de Samuel Kotz (2006). Esta obra contiene 16 volúmenes en 9,518 paginas. Para una referencia más amigable a la investigación educativa, se recomienda la obra de Salkind (2007) intitulada la Enciclopedia de Medición y Estadísticas (*Encyclopedia of Measurement and Statistics*) con tres volúmenes.

Preguntas para Reflexionar

1.- ¿Qué diferencias existen entre los análisis paramétricos y los no paramétricos?

2.- ¿Cómo se podría convertir variables continuas a frecuencias y viceversa?

3.- ¿Cuál de los análisis es más útil?

Opinión del Autor

Una manera que se puede entender la diferencia de las estadísticas paramétricas y las no paramétricas es que las primeras tienen parámetros de una población fijos y las otras no. El tener el supuesto de que hay parámetros fijos es que hay que estimarlos de algún modo como al comparar grupos por alguna variable dependiente o al correlacionar variables. Por otro lado, los análisis no-paramétricos no se asume que se va a tratar de hacer esta generalización. Solo se analizan los datos en cuanto a los datos.

Capítulo 6. Variables Dependientes e Independientes.

Un o una investigadora educativa observa que muchas de las variables de interés son constructos como el aprendizaje, la inteligencia, la autoestima, etc. Los constructos son variables que no se pueden observar a simple vista como la estatura, edad, etc. Sin embargo, se pueden manifestar de alguna manera en un test o encuesta. Un constructo como el aprendizaje, se puede definir por medio de alguna teoría: Esta sería una definición conceptual de la variable aprendizaje. ¡El problema viene una vez que se considera la forma en la que se va a medir este constructo! ¿Cómo se puede volver operacional este constructo? Una manera tradicional es el usar las calificaciones como la forma de haber vuelto operacional este constructo. Sin embargo, puede haber otras maneras de volver operacional esta variable como observar presentaciones o evaluar proyectos de estudiantes, entre otras.

Ya una vez que se tienen las definiciones conceptuales y operacionales de las variables, se puede establecer las relaciones que tienen las variables para el estudio. Por ejemplo, un o una investigadora se enfrentará la situación de definir las variables en independientes y dependientes. La excepción a lo anterior, es cuando es un análisis meramente descriptivo donde se está contestando una pregunta de investigación como: ¿Cuáles son los niveles de reprobación en las universidades mexicanas? Por el otro lado, cuando una pregunta de investigación involucre una comparación entre grupos o una relación entre variables, será el momento seguramente de precisar cuál es la variable independiente que tiene un efecto en la variable dependiente.

Antes de abordar las variables independientes y dependientes, habría que hacer un preámbulo para contextualizar esta categorización. En las preguntas de investigación se manejan variables, las cuales pasarán a las hipótesis que se verán más adelante. Por el momento, la definición de las variables es importante y estas suelen tener dos maneras: conceptual y la operacional.

Definición conceptual: Es la manera de explicar qué significa cierta variable en la literatura (cf. Hernández-Sampieri et al., 2014). Una variable que de hecho es clasificada como un constructo en la investigación educativa es el *aprendizaje* (para más información acerca de los constructos, ver a Crocker y Algina, 2006). El aprendizaje es un constructo o variable latente porque es una variable que no se puede observar y medir directamente como la estatura o la estatura de una persona. Sin embargo, se asume que el aprendizaje puede manifestarse en resolver un examen, efectuar alguna tarea. La Asociación Americana de Psicología (APA; 2015, p. 594) definió al aprendizaje como que este "…implica conscientemente o inconscientemente a asistir a los aspectos relevantes de la información entrante, organizando mentalmente la información en una representación cognitiva coherente, e integrándola con conocimiento relevante existente activado de la memoria a largo plazo." De nuevo, la anterior es una definición conceptual.

Definición Operacional: La otra definición para poder medir una variable: "Conjunto de procedimientos y actividades que se desarrollan para medir una variable" (Hernández-Sampieri et al., 2014, p. 120). Volviendo al aprendizaje, se tiene la definición conceptual, y ahora es necesario el definir como esta se va a medir (i.e., la definición operacional). Este aprendizaje se podría medir mediante exámenes y los puntajes que se obtengan en estos, entre muchas otras maneras más.

La recomendación para las investigadoras e investigadores es que revisen la literatura de acuerdo con su pregunta de investigación para dar tanto con las definiciones de conceptuales y operacionales que se han utilizado. Si no se planea utilizar una definición nueva, habría que justificarlo por medio de algún argumento. Tanto las variables dependientes o independientes tendrían definiciones conceptuales y operacionales.

En las estadísticas inferenciales, se suele nombrar a las variables como independientes (*VI*) y dependientes (*VD*). Una definición conceptual de la *VI*, según Hinkle et al. (2003, p. 736), es una variable controlada o manipulada por el investigador. Para completar las definiciones, estos últimos autores explicaron que una variable dependiente es o por lo menos se presume que es el resultado de la manipulación de la independiente. Este tipo de definición de las variables de un estudio educativo son útiles cuando se está usando preguntas de investigación de efecto:

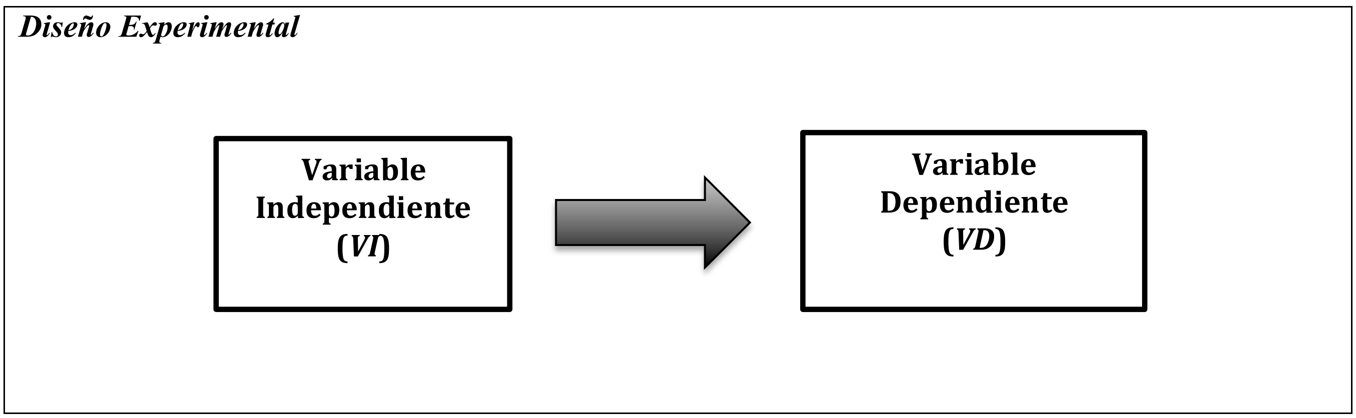
> e.g., ¿Existe un efecto de la variable independiente (e.g., horas de estudio) en la variable dependiente (calificación)?

Las variables independientes también son llamadas *predictores* y las dependientes son llamadas *criterios*. Esta terminología es común especialmente cuando se usa algún tipo de análisis de regresión. Los análisis de regresión sirven para examinar las relaciones entre los predictores y los criterios. Esta más allá del alcance de este libro el cubrir las regresiones, pero se recomienda a Tabachnick y Fidell (2012) para su uso. No siempre las variables tienen que ser divididas en independientes o dependientes. Puede pasar que el estudio solo sea descriptivo con una pregunta:

> e.g., ¿Cuáles son los niveles de calificaciones de estudiantes universitarios?

En este caso, la variable de las calificaciones no es categorizada como independiente o dependiente. Seria simplemente una variable de interés. Además, se usan variables sin definirlas como independientes o dependientes en análisis de propiedades psicométricas con el *Coeficiente del Alfa de Cronbach*, *Coeficiente del Kuder-Richardson 20 y 21*, y el *Análisis Exploratorio de Factores*. Para ampliar el tema de las propiedades psicométricas se recomienda a DeVellis (2016) quien discutió a fondo el tema del Coeficiente del Alfa de Cronbach; Kuder y Richardson (1937) crearon y explicaron sus coeficientes 20 y 21 para estimar la confiabilidad de puntajes dicotómicos (i.e, acierto o error); y Thompson (2004) dio una exhaustiva demostración de cómo llevar a cabo el Análisis Exploratorio de Factores.

Este tema de definir a una variable en *VI* o *VD* tiene también que ver con aspectos como el *diseño del estudio* (para más información acerca de diseños de estudios donde se infiere la relación de causa y efecto ver a Schneider, Carnoy, Kilpatrick, Schmidt, y Shavelson, 2007). Por ejemplo, si fuera un experimento donde hubiera dos grupos, la variable independiente seria la división de una muestra en dos grupos (idealmente tomados al azar como una muestra de una población y divididos en estos dos grupos al azar), también como haber dado el tratamiento a un grupo y un placebo al otro. La variable independiente tendría un efecto en la *VD.* Es decir, se infiere que la variable independiente manipulada por el investigador causo un cambio en la variable dependiente, sin llegar a considerar otras variables (i.e., céteris páribus; ver Figura 6.1). Estas otras variables no consideradas para la relación de causa y efecto de la *VI* y de la *VD* son las *amenazas internas* al experimento (ver a Hernández-Sampieri et al., 2014).



Nota: Bajo este diseño experimental, se asume que la manipulación de la variable independiente causo un cambio en la variable dependiente, céteris páribus.

Figura 6.1 **Variable Independiente y Variable Dependiente.**

Un o una investigadora podría echar mano de la literatura relacionada a su pregunta de investigación para definir las variables en dependientes o independientes de acuerdo con como lo han hecho otros y otras investigadoras. Esto sería algo parecido a seguir algún tipo de tradición. También, podría usar una teoría para sustentar la decisión de llamarlas *VD* o *VI*. Por ejemplo, la teoría de Piaget plantea la relación entre las etapas de la vida de una persona (i.e., *sensomotora*; *pre-operacional*; *concreta-operacional* y *formal-operacional)* y los aprendizajes que se pueden adquirir (cf. Berk, 2007, p. 20). En concreto, la variable independiente seria las etapas por las que una persona pasa al transcurrir del tiempo, y la dependiente sería los aprendizajes que se pueden adquirir. Es decir, los aprendizajes dependen de la etapa en la que la persona esta.

Otra posibilidad para definir las variables es que una variable puede ser independiente para un tipo de análisis y dependiente para otros. En la primera parte de este ejemplo heurístico de una función lineal, las calificaciones del primero de los parciales de un examen de matemáticas están en función (i.e., dependen) de las horas de estudio de cinco estudiantes. Explicando el mecanismo del ejemplo, cinco pupilos estudian diferente cantidad de horas para el primero de los exámenes parciales de matemáticas. Este parcial tiene 10 preguntas, así que obtener ninguna bien es igual al 0% de calificación y, por otro lado, obtener 10 aciertos es un 100% de calificación. La siguiente función (i.e., modelo) representa la relación entre horas de estudio como variable independiente y calificación en el primer parcial como la variable dependiente:

$y = 2x$

y = calificaciones en un parcial de matemáticas.

x = horas de estudio para el primero de los parciales de matemáticas.

Nota: En palabras, la calificación depende de la cantidad de horas de estudio multiplicadas por dos. La Tabla 6.1 muestra los resultandos cuando se efectúa esta función matemática. Los resultados evidencian a los estudiantes que más estudiaron porque obtuvieron calificaciones más altas, céteris páribus. Al no estudiar (José), la calificación es de un cero. Por otro lado, cinco horas de estudio dan como resultado un 100% de calificación para Karla.

Tabla 6.1 Horas de Estudio y las Calificaciones.

Hrs. de Estudio (x): Variable Independiente	Función y = 2x o 2x = y El resultado es el número de aciertos en el parcial	Estudiantes	Calificación en porcentaje (y) por 10 en un parcial de matemáticas: Variable Dependiente
0	2 por 0 = 0	José	0 x 10 = 0%
1	2(1) = 2	María	2 x 10 = 20%
2	2(2) = 4	Francisco	4 x 10 = 40%
3	2(3) = 6	Fernanda	6 x 10 =60%
4	2(4) = 8	Guadalupe	8 x 10 = 80%
5	2(5) = 10	Karla	10 x 10 =100%

En concreto, las horas de estudio y las calificaciones tiene una relación lineal positiva (Figura 6.2). De nuevo, cuando las horas de estudio incrementan también lo hacen las calificaciones. Este tipo de relaciones lineales son muy comunes en la investigación educativa porque son modelos simples que intentan capturar la relación entre dos variables. Sin embargo, los fenómenos educativos suelen ser más complicados que esto y habría que ver cual tipo de modelo refleja mejor la realidad. En otras palabras, los fenómenos educativos requieren en muchas ocasiones de modelos donde se usen análisis multivariados (se recomienda a Tabachnick y Fidell, 2012).

Un análisis multivariado implica el uso de más de dos variables a la vez en el modelo, y un principio epistemológico bajo el uso de este tipo de modelos es que capturen el fenómeno en la forma más integral posible. Un ejemplo de lo anterior es una relación entre el aprendizaje de un grupo de estudiantes con el nivel socioeconómico de los padres y el coeficiente intelectual de los estudiantes. En lugar de hacer dos correlaciones por separado (i.e., aprendizaje con el nivel socioeconómico y aprendizaje con el coeficiente intelectual), se haría una regresión múltiple con el aprendizaje como variable dependiente y el nivel socio-económico y el coeficiente intelectual como variables independientes. En concreto, el modelo multivariado sería preferible bajo el principio epistemológico. Asimismo, un modelo multivariado tiene otras ventajas como no inflar el Error Tipo I al llevar a cabo varios test de significancia estadística (para una discusión sobre las ventajas de los análisis multivariados, se recomienda a Tabachnick y Fidell, 2012).

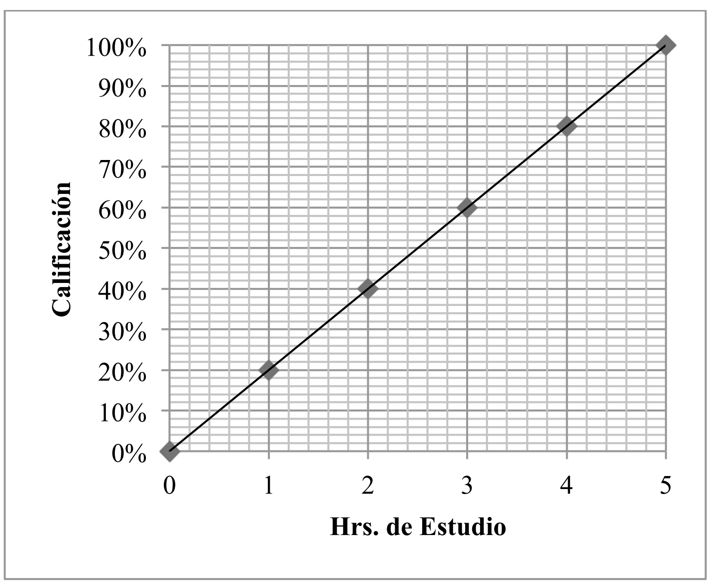


Figura 6.2 **Relación Lineal entre las Hrs. de Estudio y las Calificaciones.**

En la segunda parte del ejemplo heurístico, se cambian las definiciones de la variable independiente y dependiente para las calificaciones y las horas de estudio respectivamente. La situación ahora es que ya pasaron el primero de los exámenes parciales de matemáticas, y los y las estudiantes tienen una idea de cuánto tiempo invertir para sacarse cierta calificación en el segundo parcial. Más en detalle, los estudiantes que sacaron bajas calificaciones se sienten presionados por incrementar estas calificaciones para el segundo parcial, y los que se sacaron mayor calificación no tienen esta presión. Es decir, ahora la variable independiente es la calificación del primero de los parciales que tiene un efecto en las horas de estudio para el segundo parcial de matemáticas. Siguiendo una función similar a la anterior porque esta también es lineal, se cambian el nombre de las variables:

$y = 5 - (½ * x)$

y = horas de estudio para el segundo parcial de matemáticas.

x = calificación en el primero de los parciales de matemáticas expresado en número de aciertos (e.g., 20% = 2; 40% = 4; 60% = 6; 80% = 8; y 100% = 10).

* = signo de multiplicación

Nota: En palabras, las horas de estudio dependen de cinco unidades menos la calificación del primer parcial de matemáticas por un medio.

Tabla 6.2 Las Calificación y las Horas de Estudio.

Calificación en número de aciertos en el primer parcial de matemáticas: Variable Independiente	Función y = 5 – (½ * x)	Estudiantes	Hrs. de Estudio (y): Variable Dependiente
0	5 – (½ * 0) = 5	José	5
2	5 – (½ * 2) = 4	María	4
4	5 – (½ * 4) = 3	Francisco	3
6	5 – (½ * 6) = 2	Fernanda	2
8	5 – (½ * 8) = 1	Guadalupe	1
10	5 – (½ * 10) = 0	Karla	0

La Tabla 6.2 muestra que José quien obtuvo un cero en el primero de los parciales de matemáticas fue él que más horas de estudio invirtió: 5. Por otro lado, Karla quien había obtenido un 100% en el primero de los parciales paso cero horas estudiando. En este caso, las calificaciones y las horas de estudio tienen una relación lineal negativa (Figura 6.3). Cuando las calificaciones del primer parcial incrementan, las horas de estudio para el segundo parcial disminuyen.

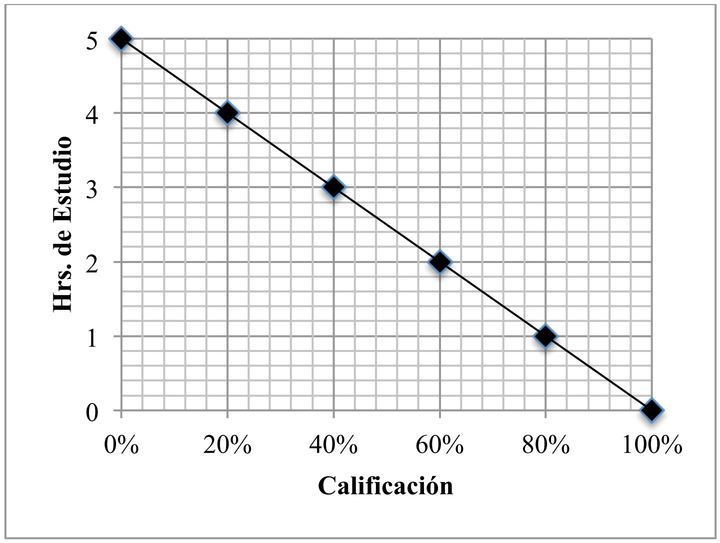


Figura 6.3 **Relación Lineal entre las Calificaciones y las Hrs. de Estudio.**

Se podría decir que existe un efecto de las horas de estudio en el primero de los parciales de matemáticas, céteris páribus. Asimismo, se podría decir que la calificación del primero de los parciales tuvo un efecto en las horas de estudio para el segundo parcial, céteris páribus. Es común hablar de efectos y tamaño del efecto en la investigación educativa. En el diccionario de Psicología del APA (*APA Dictionary of Psychology*, 2015, p. 352) se definió al efecto (*effect*) y al tamaño del efecto (*effect size*) de la misma manera, "Un evento o estado que se produce como el resultado de otro (causa)." Una interpretación de lo anterior es que, si hay un efecto entre una variable independiente y una dependiente, es que hay una relación de causa y efecto entre ambas. Esta relación causal se representa con una flecha:

Variable independiente → Variable Dependiente

Sin embargo, en la educación educativa se suele llevar a este tipo de inferencia de causa y efecto cuando se usan diseños experimentales o se usan ecuaciones estructurales. Para más información al respecto de estos temas, se recomienda ver a Schneider et al. (2007) y Byrne (2009) respectivamente. Asimismo, cuando en investigación educativa se carece de un diseño experimental y no se alude a alguna teoría para explicar la relación entre variables, se dice que hay una relación entre variables o diferencia entre grupos sin usar el término de causa y efecto. Por otro lado, aunque se carezca de diseño experimental y teoría, si se usa un tamaño de efecto como el *d* de Cohen (más adelante se explica este y para más información se recomienda a: Cohen, 1988; Cumming, 2013).

Una advertencia son las correlaciones espurias entre variables. En otras palabras, estas son correlaciones falsas que surgen simplemente por usar una fórmula de correlación como la Correlación Producto Momento de Pearson (*r*), pero carecen de un sustento teórico o de un argumento válido. Un ejemplo de una correlación espuria es cuando se relaciona el coeficiente intelectual de un grupo de estudiantes y el tamaño de sus pies. Cabe la posibilidad que aparezca una correlación de Pearson (*r*) de más de .50 por mera casualidad. Sin un sustento teórico o argumento válido, no se podría tomar esta correlación como una relación verdadera entre estas variables antes mencionadas. En otras palabras, esta última sería una correlación espuria.

Otra advertencia es cuando se usan experimentos en la investigación educativa para hacer inferencias de causalidad. Habría que tener cuidado cuando se concluye que existe una relación de causa y efecto o que no la existe porque cabe la posibilidad que otra variable haya intervenido. Esta variable que se le conoce como *amenaza interna* al experimento. Esta amenaza interna pudo haber intervenido para la obtención de resultados favorables que apoyen el tratamiento o resultados desfavorables que no apoyen el tratamiento. Un ejemplo de esto es cuando en un diseño experimental educativo donde se trata de dar tutorías de matemáticas a grupo de jóvenes para que mejoren su aprendizaje de esta materia. Una maestra se encarga de dar estas tutorías de matemáticas como el tratamiento al grupo. Antes de que ella empiece el tratamiento, se mide el conocimiento de matemáticas de los estudiantes para tener una base para ser comparada después con los resultados del tratamiento. Comienza el tratamiento y la maestra cubre el material de la forma en la que se indicó. Sin embargo, ella tiene un estilo *autoritario* que no es compatible con la manera de aprendizaje de los pupilos. El resultado es que al comparar los aprendizajes de antes y después del tratamiento no ha habido cambio. Se concluye erróneamente que el tratamiento no funciona. Por otro lado, no se consideró el estilo de la maestra el cual, según este ejemplo, intervino en la falta de aprendizaje. Esta es solo una amenaza interna a un diseño experimental. Para ahondar en las amenazas internas a los diseños experimentales, se recomienda a Hernández-Sampieri et al. (2014), y a McMillan y Schumacher (2007).

Preguntas para Reflexionar

1.- ¿Aproximadamente cuantas teorías podrían definir el constructo de aprendizaje?

2.- ¿De cuentas maneras diferentes se podría volver operacional el constructo del aprendizaje?

3.- ¿Qué otros constructos existen para el interés de la investigación educativa?

4.- ¿En qué se puede respaldar un o una investigadora para volver operacional un constructo?

5.- ¿De qué depende de que una variable independiente o dependiente hayan sido definidas como tales?

6.- ¿Una vez que se definió a una variable como independiente debe de usarse como tal en cada uno de los estudios subsecuentes?

7.- ¿Cuándo se le considera una relación de causa y efecto entre la variable independiente o dependiente?

8.- ¿Qué amenazas internas pueden interponerse en hacer de causa y efecto entre las variables?

Opinión del Autor

Mis estudiantes se quedan en muchas veces perplejos o perplejas cuando tienen que diferenciar entre un constructo y la forma de medirlo: definición conceptual y operacional. Posiblemente sea porque en forma intuitiva uno piensa que estos son sinónimos. Esto, tal vez, se puede deber a uno como estudiante crece en ambientes escolares donde la inteligencia o el esfuerzo son sinónimos de buenas calificaciones. En otras palabras, si una estudiante logra excelencia académica o es porque es muy inteligente o se esforzó más que el resto. También, se considera que es una combinación de ambos elementos. Más allá del efecto que pueden tener la inteligencia y el esfuerzo en las calificaciones, difícilmente se considera a el aprendizaje como la manifestación de estas últimas, céteris páribus. Aquí es cuando las definiciones conceptuales y operacionales pueden cobrar sentido al establecer que el aprendizaje se manifiesta en forma de las calificaciones.

Otra manera de quedarse perplejo o perpleja es cuando se tiene que definir a las variables en independientes o dependientes. Para algunos de mis estudiantes, el establecimiento de las variables como independientes o dependientes es parecido al dilema del huevo y la gallina: i.e., ¿Cuál fue primero? Más allá de tratar de solucionar esta controversia que implicaría probablemente usar la Teoría de la Evolución de Carlos Darwin, la respuesta a la definición de variables es más simple. Se trataría de establecer que sucedió primero y el efecto que tiene sobre otras cosas. Por ejemplo, el desarrollo físico y psicológico de una persona desde su concepción hasta su muerte podrían ser variables independientes de los conocimientos que adquieren durante su vida. Por otro lado, estos conocimientos pueden tener un efecto en su estilo de vida como dieta su nivel educativo, filosofía, etc. ¡Esto se empieza a complicar porque como dicen muchos teóricos: todo está relacionado con todo! Pasa que un investigador o una investigadora solo está estudiando una pequeña parte de la realidad mediante un modelo matemático, el cual le puede permitir establecer diferencias o asociaciones entre variables de las que se pueden inferir relaciones de causa y efecto. Lo anterior suena bastante abstracto, pero es mi intención el tratar de capturar lo más que se pueda de las variables independientes y dependientes al igual que su relación causal. También, trato de decir que probablemente no hay absolutos en cuestión de definiciones de *VI* y *VD* para todos los estudios. En concreto, mi recomendación es basarse para definirlas en cuestión de establecer lógicamente cual tendría un efecto de una sobre la otro, usar la tradición de sus definiciones o modelos teóricos.

Capítulo 7. Hipótesis Nula y Alternativa.

En programas o películas donde hay un juicio, se tiene una parte acusadora (i.e., un o una fiscal) y la parte defensora (i.e., el abogado o abogada). En estas situaciones, se maneja mucho el termino de ser inocente por parte del presunto criminal. Si el programa o película se apegan un poco a la terminología legal, no usan el termino de inocencia. La razón es que probar la inocencia es muy complicado. ¿Es decir, como se podría mostrar que una persona no tuvo que ver en un crimen? Por ello, se opta por tener los términos de culpable o no culpable. De hecho, se parte de la idea de que una persona es no culpable y el peso de la evidencia para declararla culpable recae en la fiscalía. Es decir, un acusado no tiene que demostrar que es inocente, sino que él o la fiscal debe mostrar la culpabilidad. Por ejemplo y en un asesinato, si la evidencia está muy relacionada a las acciones o situación de un acusado como el tener un móvil (e.g., quedarse con una prima de seguro de vida), los medios (e.g., poseer armas de fuego) y la oportunidad para hacerlo (e.g., se encontraban cazando juntos la víctima y su supuesto verdugo), entonces lo podrían encontrar culpable por parte de un juez o jurado. Por otro lado, si no hay evidencia y móvil que liguen al presunto asesino a la muerte de una persona, probablemente lo declaren no culpable. Ambos veredictos de culpabilidad o no-culpabilidad son por lo regular inferencias que se hacen a partir de la evidencia.

De un modo similar en la investigación, se hacen inferencias a partir de las evidencias que se tienen de una porción de personas, observaciones u objetos. Por ejemplo, se va de una muestra de estudiantes con un índice de reprobación del 10% a generalizar que una población de donde se sacó la muestra también tiene aproximadamente un índice de 10% de reprobación. También como en los juicios, se parte de una hipótesis nula, en la cual se estipula que no hay diferencia entre los grupos que se están comparando o relación entre variables. La hipótesis nula de los juicios es se asume que el sospechoso no es culpable porque no hay relación entre sus acciones o situación con el crimen. Por lo menos, se parte de este supuesto. La investigadora es la análoga a la fiscal que debe de mostrar la relación entre la evidencia y la conclusión de que hay una diferencia entre grupos o relación entre variables para poder rechazar la hipótesis nula. Al rechazar la hipótesis nula basado en los resultados arrojados por análisis estadístico, se tiene evidencia de que hay un efecto estadístico entre una variable y otra. La diferencia entre la investigadora y la fiscal es que esta primera no tiene una posición a ultranza como la fiscal. Es decir, puede pasar que los resultados de los análisis muestren evidencia que no se puede rechazar la hipótesis nula. En este caso, el o la investigadora acepta que no se puede rechazar la hipótesis nula y que aparentemente no hay un efecto estadístico entre la variable independiente y dependiente.

Al igual que en los veredictos de los juicios, en la investigación educativa se pueden cometer errores. Por ejemplo, un veredicto puede ser que una persona es culpable cuando realmente él no lo hizo. De una manera similar, se puede rechazar que una hipótesis nula, usando una muestra, y concluir que hay un efecto estadístico entre las variables cuando en la población no hay un efecto. También, puede pasar en un juicio que una persona sea declarada no culpable cuando realmente cometió el crimen. En las estadísticas el paralelo de lo anterior seria que no rechazo la hipótesis nula en base a los resultados de la muestra cuando en las poblaciones había un efecto entre las variables. Este último párrafo puede sonar complicado porque faltan de explicar algunos elementos entre poblaciones y muestras, pero se trataran de explicar estos elementos durante este capítulo.

Existen varias teorías de la probabilidad (para más información de las teorías de la probabilidad clásica y moderna, ver a Fischer, 2010). La parte de la probabilidad de este presente libro está dentro de lo que se denomina la *Teoría Clásica de la Probabilidad* (*Classical Probability Theory*).

En el siguiente párrafo, se recapitulan algunos de los conceptos que se habían cubierto al principio en este Libro. En las estadísticas inferenciales, una vez obtenida una muestra se trata de hacer *inferencias* acerca de una población: Se va de las estadísticas de la muestra a estimar los parámetros de la población. Por ejemplo, si al sacar el promedio ($\bar{x}$) de una muestra se obtuvo una calificación en matemáticas de 87.5 puntos con un intervalo de confianza del 95%, se asume que, de 100 muestras similares a la primera, 95 de ellas capturan el promedio de la población (μ). Esta suposición también descansa en el Teorema de tendencia Central porque esa muestra con su promedio probablemente caería en el centro de la *distribución de muestras*, capturando el promedio de la población. Hablando de muestras, Schumacker y Tomek (2013) explicaron que una población es un set bien definido de personas, eventos u objetos, de la cual se obtiene una muestra representativa que es un sub-set de individuos, eventos u objetos. Para tener una oportunidad de tener una muestra representativa, cada uno de los integrantes debe de tener la misma oportunidad de ser seleccionado al azar. También, se asume que hay un error de muestreo que es la diferencia entre el promedio de la población (μ) y el de muestra ($\bar{x}$):

Error de muestreo del promedio = μ - $\bar{x}$

En las estadísticas inferenciales, se derivan hipótesis de las preguntas de investigación y se ponen a prueba mediante un análisis estadístico. En otras palabras, estas se ponen a prueba para ver si los datos las apoyan. De este modo y con un ejemplo de una medición de aprendizaje, se parte de una hipótesis nula (H_0), la cual implica que no hay diferencia entre grupos en lo que respecta a la variable dependiente (e.g., *VD* = calificaciones). Con respecto a correlaciones, la hipótesis nula también especifica que no existe relación entre dos variables (e.g., una *VI* y una *VD* en una correlación). Usando esta hipótesis nula, se asume que no hay un efecto de la variable independiente en la dependiente. Por otro lado, se tiene la hipótesis alternativa (H_A) que implica una diferencia entre grupos o relación entre variables. En corto, la hipótesis alternativa implica un efecto ($VI \rightarrow VD$). Hipotéticamente hablando, hay un estudio donde se están comparando dos grupos (niñas contra niños) en cuestión de puntajes en un examen de matemáticas bajo la pregunta de investigación:

¿Existe un efecto del sexo en los puntajes de un examen de matemáticas?

Esta misma pregunta se puede expresar de otro modo sin que pierda su significado para hacer más simple la derivación de las hipótesis:

¿Existe una diferencia entre el promedio de los puntajes de las niñas y de los niños?

Estas preguntas corresponderían mejor a un diseño no-experimental porque el investigador no controla el tratamiento ya que este es el sexo (para más información acerca de los diseños experimentales y no-experimental, se recomienda a Hernández-Sampieri et al., 2014, y a McMillan y Schumacher, 2007). Sin embargo, es común comparar grupos, en los diseños experimentales, en cuanto a una variable dependiente (ver a Maxwell y Delaney, 2004 para análisis complejos de análisis de la varianza). Las hipótesis se establecen desde la perspectiva que se están comparando dos poblaciones diferentes. Es decir, los niños vienen de una población y las niñas de otra diferente. Por ello, se expresa el promedio de las calificaciones de matemáticas con la letra griega (μ = promedio de una población que se pronuncia: *mu*). Entonces, las hipótesis nula y alternativa quedan de la siguiente manera:

Ejemplo de hipótesis nula (H_0): $\mu_{\textit{niñas}} = \mu_{\textit{niños}}$

Los promedios de las dos poblaciones son iguales.

Otra manera de expresar lo anterior es H_0: $\mu_{\textit{niñas}} - \mu_{\textit{niños}} = 0$

Al restar ambos promedios de la población da como resultado un cero porque son iguales.

Para las muestras se usa la siguiente nomenclatura:

hipótesis nula (H_0): $\bar{x}_{\textit{niñas}} = \bar{x}_{\textit{niños}}$

Los promedios de las dos muestras son iguales.

Otra manera de expresar lo anterior es H_0: $\bar{x}_{\textit{niñas}} - \bar{x}_{\textit{niños}} = 0$

Al restar ambos promedios de las muestras dan como resultado un cero porque eran iguales.

Ejemplo de hipótesis alternativa (H_A): $\mu_{\textit{niñas}} \neq \mu_{\textit{niños}}$

Los promedios de las dos poblaciones *no* son iguales.

Otra manera de expresar lo anterior es: H_A: $\mu_{\textit{niñas}} - \mu_{\textit{niños}} \neq 0$

Al restar ambos promedios de las poblaciones *no* dan como resultado un cero porque eran diferentes.

Asimismo, para las muestras:

hipótesis alternativa (H_A): $\bar{x}_{niñas} \neq \bar{x}_{niños}$

Los promedios de las dos muestras *no* son iguales.

Otra manera de expresar lo anterior es: H_A: $\bar{x}_{niñas} - \bar{x}_{niños} \neq 0$

Al restar ambos promedios de las muestras *no* dan como resultado un cero porque eran diferentes.

En esta serie de hipótesis alternativas, no se especificó si un promedio de un grupo era mayor que el otro. Para este tipo de hipótesis se usa una curva normal con *dos colas* que contiene las diferencias entre los promedios. Más adelante, se precisará este concepto de un análisis con dos colas.

Las hipótesis alternativas pueden tener un sentido: una cola que se explicara más adelante también. Una hipótesis alternativa con sentido se especifica si el promedio uno será más grande que el promedio dos y viceversa. Por ejemplo, si se cree que el promedio de la población de niñas será más grande que la población de niños se puede estipular de la siguiente manera:

Ejemplo de hipótesis alternativa con sentido (H_A): $\mu_{niñas} > \mu_{niños}$

El promedio de la población de niñas será más grande que el de los niños.

Asimismo, para las muestras:

hipótesis alternativa (H_A): $\bar{x}_{niñas} > \bar{x}_{niños}$

El promedio de la muestra de niñas será más grande que el de los niños.

La Figura 7.1 muestra las dos distribuciones normales que representan los datos de las poblaciones de niñas y niños. En este caso, la distribución de las niñas está más hacia la derecha (+∞) que la de los niños. Esto indica que el promedio de las niñas es mayor que el de los niños. Sin embargo, habría que practicarle un test de significancia estadística para estimar si hay una diferencia significativa entre ambos promedios. Más adelante en el presente libro, se efectuarán *Test de Significancia Estadística de la Hipótesis Nula* (*Null Hypothesis Significance Testing*; ver a Cumming, 2013, para más detalles).

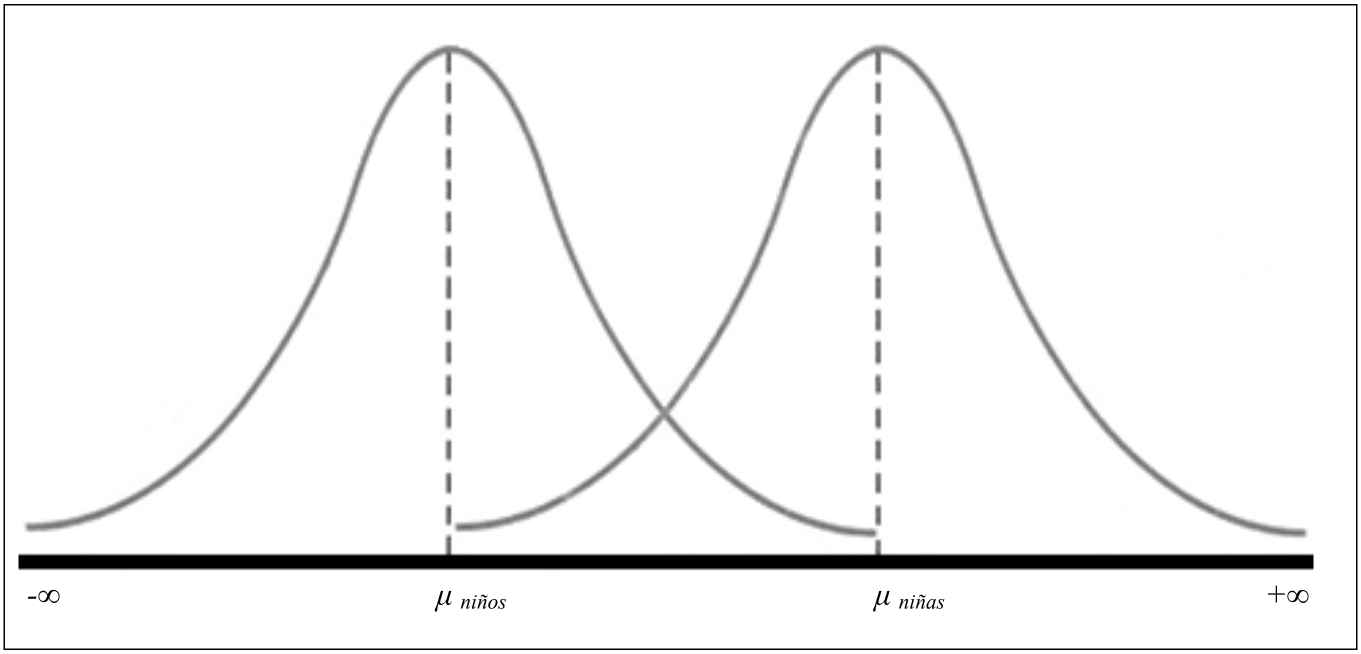


Figura 7.1 **Dos Distribuciones Normales para los Promedios de los Niños y las Niñas.**

Una hipótesis nula se puede *rechazar* o *fallar en rechazar* (i.e., estas son las dos posibilidades que existen en este caso, y a estas dos posibles alternativas se les llama *muestra espacial*, *simple space*; ver a Ross, 1997 para más detalles). *Fallar en rechazar* significa: no se rechazó. Esta más allá de los propósitos del presente libro el explicar este vocabulario estadístico matemático, pero si se quiere más información se recomienda a Salkind (2007). Una sola curva normal puede ser usada para representar la diferencia que puede haber al comparar dos promedios. La construcción de esta curva normal que representa las diferencias entre promedios será explicada más adelante. Por el momento, se toma esta curva normal para representar las diferencias entre promedios (ver Figura 7.2):

Área sin estar sombreada: Es el área de *no rechazo* (i.e., fallar en rechazar) de la hipótesis nula para un test de dos colas (área entre el 0 y un valor critico dado por α / 2), lo cual implica:

H_0: $\mu_{niñas} = \mu_{niños}$.

Área sombreada: Es el área de rechazo a la hipótesis nula para un test de dos colas (área que está más allá del valor critico dado por α / 2), lo cual implica:

H_0: $\mu_{\, niñas} \neq \mu_{\, niños}$.

Nota: Un valor critico está en el mismo punto que el alfa dividido entre dos que sirve de límite para establecer las áreas de rechazo y no rechazo de la hipótesis nula (más detalles al respecto se darán más adelante).

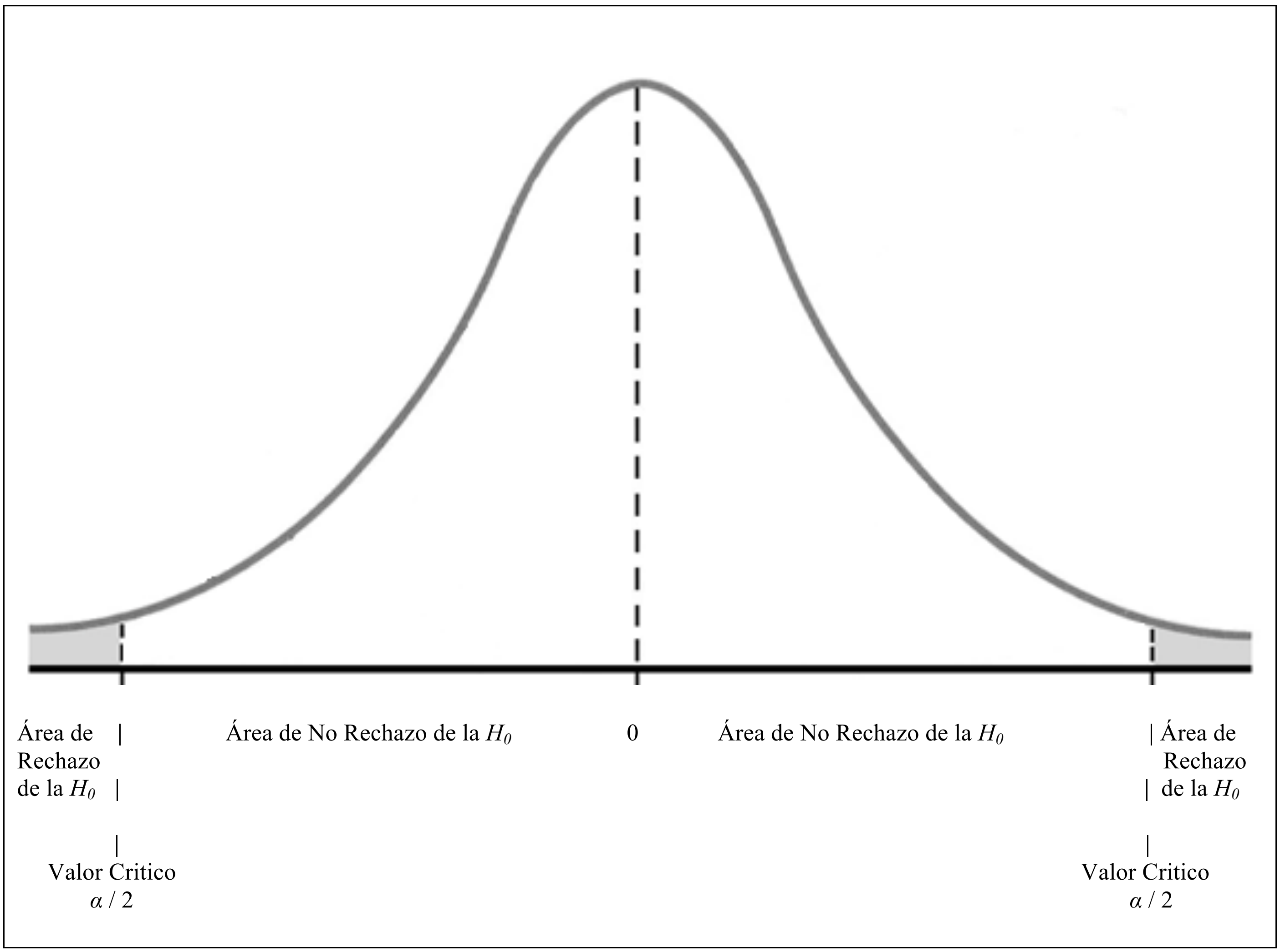


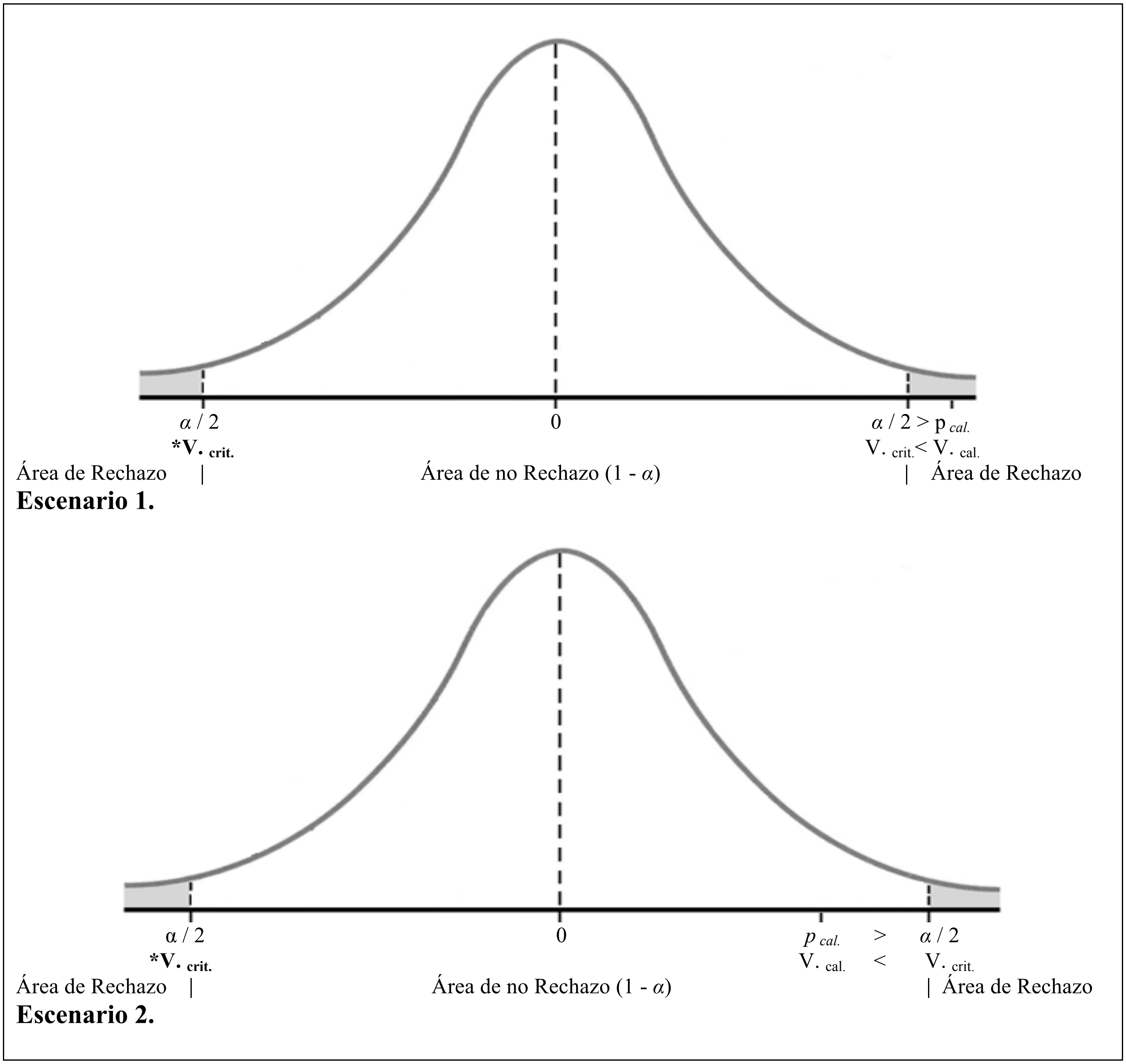
Figura 7.2 **Áreas de Rechazo y no Rechazo de la Hipótesis Nula.**

Para complementar, la Tabla 7.1 muestra una serie de pasos para saber si se rechaza o se falla en rechazar una hipótesis nula (H_0):

Tabla 7.1 Poniendo a Prueba una Hipótesis: Procedimiento de Test de Significancia Estadística para la Hipótesis Nula.

Paso 1	Derivar las hipótesis nula y alternativa de la pregunta de investigación. E.g. utilizando la pregunta: ¿Existe una diferencia entre el promedio de los puntajes de las niñas y de los niños? Entonces, las hipótesis son: H_0: $\mu_{\,niñas} = \mu_{\,niños}$ H_A: $\mu_{\,niñas} \neq \mu_{\,niños}$
Paso 2	Seleccionar algún tipo de análisis estadístico: e.g., para el caso anterior un *test t* de estudiante o análisis de la varianza (ANOVA) serían factibles.
Paso 3	Identificar el nivel de significancia que se desee. También, se le llama el nivel del alfa (α) e implica un valor crítico. Este nivel de significancia (α) ha sido tradicionalmente expresado en un porcentaje y su correspondiente decimal: 1% (.01), 5% (.05) y 10% (.10). Este porcentaje (área bajo la curva) va desde un punto crítico hacia el infinito. El área más allá del nivel de significancia se le llama el área de rechazar la hipótesis nula. El porcentaje que queda (1 - α) representa un área bajo la curva normal que va desde un punto crítico hasta el otro cuando se tienen dos colas o desde un punto crítico hasta el infinito cuando es de una sola cola. Esta área se le conoce como el área de no rechazar la hipótesis nula. Un ejemplo del tamaño del área (porcentaje) de no rechazo es cuando se elige un α = .01, y, por lo tanto, se obtiene: 1 - .01 = .99 que equivale al 99% del área de no rechazo. En las investigaciones educativas usualmente se usa el nivel de significancia del 5% (.05). También, al nivel de significancia (α) se le conoce como el Error Tipo I (*Type I Error*), el cual se discutirá más adelante.
Paso 4	Con el nivel de significancia (α) elegido se va a una tabla de valores críticos (ver Apéndice). Los valores críticos empiezan en cero en el promedio (e.g., valores t o z; esto no aplica a la distribución F porque no es simétrica). Del cero hacia el infinito positivo, los valores críticos van incrementando: e.g., 1, 2, 3, …, +∞. Del cero al infinito negativo, los valores críticos van disminuyendo: -1, -2, -3, …, -∞. Este valor también servirá de límite para rechazar o no rechazar la H_0.
Paso 5	Calcular de los datos los valores observados: e.g., un valor observado de la distribución $t_{\,calculado}$ o $F_{\,calculado}$. Si se usa SPSS o Excel, entre otros programas, se calculará una probabilidad observada: $p_{\,calculado}$. Estos valores calculados de los datos servirán para compararlos con los valores críticos y el nivel de significancia.
Paso 6	Comparar los valores críticos y los calculados. Estos son los dos posibles escenarios (i.e., *muestra espacial*; ver la Figura 7.3): Escenario 1: $\alpha / 2 > p_{\,calculada}$ lo cual significa que el valor $_{critico}$ < valor $_{calculado}$. En este escenario se rechaza la hipótesis nula porque la probabilidad calculada es menor al nivel de significancia, así como el valor calculado es mayor que el valor crítico. Lo anterior es bajo el supuesto de que la hipótesis nula es cierta: los promedios de las poblaciones son diferentes. Escenario 2: $\alpha / 2 \leq p_{\,calculada}$ lo cual significa que valor $_{critico}$ > valor $_{calculado}$. En el segundo escenario se falla en rechazar la hipótesis nula porque la probabilidad calculada es mayor al nivel de significancia, así como el valor calculado es menor que el valor $_{crítico}$. Lo anterior es bajo el supuesto de que la hipótesis nula es cierta: los promedios de las poblaciones son diferentes. *Nota:* Es suficiente con que la probabilidad calculada sea igual al alfa para que se falle en rechazar cuando se usan las dos colas de la distribución de la diferencia entre promedios: i.e., $\alpha / 2 = p_{\,calculado.}$ Cuando se usa una sola cola: Más adelante se verá el test *Test de Significancia Estadística de la Hipótesis Nula para una sola cola.*

Nota: El poner a prueba la hipótesis como en el cuadro anterior se le conoce también como Procedimiento de Test de Significancia Estadística para la Hipótesis Nula (cf. Cumming, 2013).



Nota: las abreviaciones son: p_{cal} = *probabilidad* $_{calculado}$; $V._{crit}$ = Valor $_{critico}$; $V._{cal.}$ = Valor $_{crítico}$. *Bajo un test de dos colas, si se elige la cola de la izquierda y se rechaza la hipótesis nula, el Valor critico será menor al Valor Calculado, en este caso, porque ambos tienen signos negativos y entre más se alejen del promedio serán más y más pequeños. Para rechazar la hipótesis nula en esta misma situación, la probabilidad calculada será menor al alfa porque el área no tiene signo negativo.

Figura 7.3 **Escenarios 1 y 2 para Rechazar o no la Hipótesis Nula.**

Después de comparar los promedios de ambos grupos, hay una diferencia entre ambos (i.e., el promedio de las niñas > el promedio de los niños). La pregunta que surge seria:

¿Ocurrió la diferencia por casualidad?

También, el procedimiento llamado Test de Significancia Estadística para la Hipótesis Nula (*Null Hypotesis Significanse Testing*) o simplemente test de significancia estadística ayuda a contestar esa interrogante. Este procedimiento estadístico es para poner a prueba la casualidad como explicación factible de un evento (Lane et al., 2014). Un ejemplo de casualidad es afirmar que el número uno va a aparecer después de tirar un dado y que esto resulte atinado: aparece. Suponiendo que no había manera de saber el resultado esto fue obra de la casualidad. Los Test de Significancia Estadística para la Hipótesis Nula han sido diseñados para minimizar la posibilidad de casualidad (cf. Hinkle et al., 2003). Test de Significancia Estadística para la Hipótesis Nula ha sido definida como (APA, 2015, p. 980):

> Un set de procedimientos usados para determinar si las diferencias entre dos grupos o modelos es estadísticamente significativa (i.e., improbablemente de surgir por casualidad). En su forma más común, un test de significancia es usado para decidir si la hipótesis nula debe de ser rechazada.

La postura inicial, de un investigador o investigadora es la hipótesis nula: i.e., no hay diferencia entre los promedios. En otras palabras, aunque entre los promedios se observe que no son exactamente iguales, se asume que esta diferencia se debió al azar cuando se parte de la H_0. Volviendo a la Tabla 7.1, la diferencia entre el promedio de los niños y las niñas fue estadísticamente significativa para el Escenario 1:

i.e., $\alpha / 2 > p_{calculada}$ lo cual significa que el valor $_{\text{critico}}$ < el valor $_{\text{calculado}}$.

Esta diferencia estadísticamente significativa se refiere a que la diferencia entre los promedios no se debió posiblemente a la casualidad, bajo el supuesto de que la hipótesis nula es cierta: no hay diferencia entre los promedios de las poblaciones. No se podría afirmar que definitivamente no se debió al azar porque se pueden cometer errores al tratar de generalizar los resultados de la muestra a los de la población (i.e., Error Tipo I y II como se verá más adelante). En otras palabras, cabe la posibilidad que al comparar los promedios de las muestras exista una diferencia estadísticamente significativa, pero al comparar las poblaciones de las que proceden estas muestras, se encuentre que entre estas poblaciones no existe una diferencia estadísticamente significativa:

Comparación de muestras: $\bar{x}_1 \neq \bar{x}_2$ (diferencia estadísticamente significativa).

Comparación de poblaciones: $\mu_1 = \mu_2$ (sin diferencia y este es el supuesto de que la nula es cierta).

Una probabilidad $_{\text{calculada}}$ baja relativamente al alfa que se eligió hace que surjan dudas acerca de la factibilidad de la hipótesis nula (cf. Lane et al., 2014). Por mera convención, se ha adoptado en muchos casos de investigación educativa el concluir en rechazar una hipótesis nula cuando la probabilidad de un evento (i.e., una diferencia; $p_{\ calculado}$) es menos del 5% (i.e., $\alpha = .05$). Esta convención del 5% del alfa y el origen de la Test de Significancia Estadística para la Hipótesis Nula son cubiertas por Yu (2006). Cuando la hipótesis nula es rechazada, el efecto estadístico se dice que es estadísticamente significativo. Esto no significa necesariamente que el efecto estadístico es importante porque no se está midiendo su magnitud por el momento. Por ejemplo, la magnitud del efecto (*effect size*) también llamado *efecto práctico* se podría medir con el *d* de Cohen (ver a Cohen, 1988). Este Test de Significancia Estadística para la Hipótesis Nula es sensible al tamaño de la muestra, así que si se tiene una muestra lo suficientemente grande se podría tener un efecto estadísticamente significativo (cf. Hinkle et al., 2003; Lane et al., 2014; Salkind, 2007; Tabachnick y Fidell, 2012). Por otro lado, el efecto práctico no es sensible al tamaño de la muestra en su fórmula (esta fórmula se cubrirá en el Capítulo 11). En resumen, dos tipos de efectos se pueden calcular cuando se comparan los promedios de dos muestras:

Efecto estadístico: Test de Significancia Estadística para la Hipótesis Nula.

Efecto práctico: e.g., *d* de Cohen (este efecto se verá más adelante), entre otros efectos.

De hecho, Cumming (2013) ha sido uno de los defensores de darle más importancia al efecto práctico que al estadístico. La razón es porque el efecto práctico provee la posibilidad de apreciar la magnitud de la diferencia. Esto no sucede con el efecto estadístico.

Volviendo al contexto de la significancia estadística (Test de Significancia Estadística para la Hipótesis Nula), un test de significancia es llevado a cabo y la probabilidad $_{\text{calculada}}$ refleja la fuerza de la evidencia en contra de la hipótesis nula (Lane et al., 2014). Por ejemplo y en un test de dos colas, si la probabilidad $_{\text{calculada}}$ es menor a un alfa de .05 / 2 = .025 por cola, los datos proveen evidencia fuerte de que la hipótesis nula es falsa. Si la probabilidad $_{calculada}$ esta entre .01 y menor a .05, la nula es típicamente rechazada, pero no con tanta confianza como si hubiera estado debajo de .01 (Lane et al., 2014). Estos últimos autores también añadieron que las probabilidades calculadas entre .05 y .10 proveen evidencia débil en contra de la hipótesis nula y, por convención, no son lo suficientemente pequeñas como para rechazar la hipótesis nula. Probabilidades aún más altas a .10 proveen menos evidencia que la hipótesis nula es falsa. No se está calculando la probabilidad de que la hipótesis es falsa para ello están las estadísticas Bayesianas que proveen métodos específicos para computar las probabilidades de las hipótesis (las probabilidades requieren que se especifique la probabilidad de las hipótesis antes de que los datos sean considerados, y, por ello, es difícil de ser implementado en otros contextos). Para usar análisis probabilísticos Bayesianos, se recomienda ver a Box y Tiao (1973).

Preguntas para Reflexionar

1.- ¿Por qué se toma de referencia la hipótesis nula?

2.- ¿Puede suceder que la probabilidad calculada sea menor al alfa mientras el valor calculado sea menor al valor critico?

3.- ¿Qué se debería hacer si la hipótesis nula no es rechazada, pero existe un cumulo de literatura donde la hipótesis nula ha sido rechazada?

4.- ¿Cómo suele interpretar el o la investigadora promedio el rechazo de la hipótesis nula?

5.- ¿Qué tanto aparece el efecto estadístico más que el efecto práctico?

6.- ¿A qué se debe que el efecto estadístico es más comúnmente usado que el efecto práctico?

7.- ¿Cómo se podría justificar el usar un alfa diferente al alfa = .05?

8.- ¿Qué otras teorías probabilísticas se pueden usar para calcular las probabilidades en el contexto de Test de Significancia Estadística para la Hipótesis Nula?

9.- ¿Cómo se podrían usar estas teorías de probabilidad?

Opinión del Autor

Para mí, la hipótesis nula refleja la postura de escepticismo de la que parte una investigadora o un investigador. En concreto, la hipótesis nula afirma que no hay una diferencia entre grupos en cuestión de la medición de alguna variable dependiente. Un ejemplo sería el de la inteligencia como variable dependiente de la nacionalidad de ciertos grupos. Es decir, una nacionalidad representaría a un grupo y todos tendrían el mismo nivel de inteligencia. Entonces, se pondría a prueba esta hipótesis nula para observar si la evidencia muestra un efecto estadístico por parte de la nacionalidad en la inteligencia ($\text{probabilidad}_{\text{calculada}} < \text{alfa}$). También, la hipótesis nula se pone a prueba para observar si algunas variables están relacionadas entre sí o no. La hipótesis nula es que no lo están. Un ejemplo puede ser entre la edad y la inteligencia. La nula indica que la edad como variable independiente no tiene efecto en la inteligencia ($\text{probabilidad}_{\text{calculada}} > \text{alfa}$). En otras palabras, no hay efecto de la edad en la inteligencia. Puede pasar que si tenga un efecto estadístico ($\text{probabilidad}_{\text{calculada}} < \text{alfa}$).

Una hipótesis nula puede ser rechaza o no. Esta acción de no rechazar también se le llama: *Fallar en rechazar*. Lo que no se dice es que se acepta la hipótesis nula. Esta más allá del alcance de este libro el tratar de explicarlo. Sin embargo, existen una serie de enciclopedias estadísticas que discuten lo anterior, pero también hay muchos libros de estadísticas que insisten en decir que se acepta la hipótesis nula. Habría que tener cuidado y no seguir lo que estos últimos libros explican.

El efecto estadístico es para comparar la $\text{probabilidad}_{\text{calculada}}$ de algún test estadístico contra un criterio como es el alfa. Al ser más grande una $\text{probabilidad}_{\text{calculada}}$ que el alfa, se puede concluir que la diferencia que se estaba estudiando entre dos promedios fue probablemente al azar. Sería como arrojar una moneda al aire en 100 ocasiones y que de estos lances: 51 fueran cara y 49 fueran cruz. Es decir, fue una diferencia al azar entre estas frecuencias. Si se volvieran a arrojar, podría ser 50 para la cara y 50 para la cruz. Esto es lo que se esperaría en probabilidad clásica: 50% y 50%. Por otro lado, si la cara solo apareciera en 4 ocasiones de 100 (i.e., .04) y el criterio fuera un alfa = .05, habría una diferencia significativa y se rechazaría la hipótesis nula. Se concluiría que existe un efecto de la moneda en las frecuencias.

Algo que no mide los test de significancia estadística es la magnitud de la diferencia entre dos promedios. Mucha gente cree que si obtiene una probabilidad $_{\text{calculada}}$ = .0001 es mejor que una probabilidad calculada = .04 porque esta primera es mucho más pequeña que el tradicional coeficiente alfa = .05. Autores como Lane et al. (2014) han categorizado los diferentes niveles de alfa como evidencia a favor de una hipótesis nula falsa. Es decir, si una probabilidad $_{\text{calculada}}$ se acerca al alfa, se considera que provee menos evidencia de que la nula es falsa. Al contrario, una probabilidad $_{\text{calculada}}$ que se aleja del alfa provee evidencia de que la nula es falsa. Lo anterior también lo defienden otros autores: una probabilidad $_{\text{calculada}}$ menor pero cercana al alfa, se debe de tratar con reservas antes de rechazar la nula. Aunque se considere la magnitud de una (probabilidad $_{\text{calculada}}$), esto sigue sin medir la diferencia entre los promedios. Habría que usar un efecto práctico para calcular el tamaño de la diferencia como el *d* de Cohen.

Capítulo 8. Error Tipo I y II.

Como se explicó en el preámbulo de la sección anterior: Cuando se da un veredicto en un juicio, este puede ser erróneo. Esto es, un enjuiciado puede ser declarado culpable cuando no cometió el crimen. Por otro lado, puede ser declarado no culpable cuando en realidad cometió el crimen. De una manera similar en la investigación educativa, se pueden sacar inferencias que no corresponden a la realidad como en un juicio. Una de estas inferencias erróneas es cuando se tienen dos muestras de personas de dos diferentes poblaciones y se concluye que hay una diferencia en las muestras. Sin embargo, en las poblaciones de origen de estas poblaciones, no había diferencia. En otras palabras, cuando se hace la generalización de los resultados de la muestra a la población, se cae en el error. Otro error es cuando se encuentra que no hay diferencia entre dos muestras, pero si la había en cuanto a las poblaciones. Estos son los dos tipos de errores que se pueden cometer al trabajar con muestras y hacer generalizaciones que no corresponden a la realidad de las poblaciones.

Al rechazar o fallar en rechazar una hipótesis nula, se pueden cometer dos tipos de errores. Los Errores Tipo I y II se cometen cuando hay una discrepancia entre comparaciones de unas muestras (e.g., comparación entre los promedios de estas) y las comparaciones de los promedios de sus respectivas poblaciones. En otras palabras, una comparación entre los promedios de unas muestras de diferentes poblaciones resulta en que hay una diferencia entre ambos promedios, pero al comparar los promedios de las poblaciones resulta en que no hay diferencia. Entonces, lo anterior es un error porque se encontró una diferencia entre promedios de las muestras cuando no la había entre las poblaciones. Asimismo, otro error sería cuando no se encuentra diferencia entre los promedios de las muestras, pero si la hay entre los promedios de las poblaciones. En forma breve, las definiciones de los Errores Tipo I y II fueron dadas por Hinkle et al. (2003, p. 740):

Error Tipo I (*Type I Error*, α): Es rechazar la hipótesis nula cuando era de hecho *cierta.*

Error Tipo II (*Type II Error,* β): Es fallar en rechazar la hipótesis nula cuando de hecho es *falsa.*

Más en detalle, *la probabilidad del Error Tipo I* es llamada el nivel de significancia del test (i.e., α). Por otro lado, Lane et al. (2014) explicaron que el alfa no es exactamente el Error Tipo I. En lugar de lo anterior, el alfa es la probabilidad del Error Tipo I, asumiendo que la nula sea cierta (Lane et al., 2014). La hipótesis nula es puesta a prueba a nivel de significancia (α). En la investigación en general, se maneja un nivel del alfa de .05 o menos. Las razones para esta costumbre no son del todo claras, pero usualmente se le atribuye a Fisher (ver a Kotz, 2006). El nivel de significancia (α) es elegido por la investigadora o investigador. El Error Tipo I es afectado por alfa elegido (nivel de significancia): más bajos niveles del alfa implican una incidencia menor del Error Tipo I. Además del alfa = .05, también existen coeficientes alfa de .001, .01, o .10. En otras palabras, si se va de un coeficiente alfa de .10 a uno de .05, se disminuye la probabilidad del Error Tipo I.

Un ejemplo heurístico del Error Tipo I es que se comparan los promedios de matemáticas de una muestra de niños contra una de niñas y resulta una probabilidad $_{\text{calculada}}$ de .001 cuando se está comparando a un alfa de dos colas de .05 / 2. Es una diferencia estadísticamente significativa e implica el rechazar la hipótesis nula. Sin embargo, estas dos muestras vienen de dos poblaciones que no necesariamente son diferentes en sus promedios: la de las niñas y la de los niños. Continuando y bajo el supuesto de que no hay diferencia entre las poblaciones, se estaría cometiendo un error al rechazar la nula porque la hipótesis nula era cierta para la población. Todo esto es bajo la premisa de que son estadísticas inferenciales y se quieren generalizar los resultados de las muestras a las poblaciones. Con más precisión, el error se llama de Tipo I que sucede cuando se rechaza una hipótesis nula que es cierta.

Comparación de muestras: $\bar{x}_{\text{niñas}} \neq \bar{x}_{\text{niños}}$.

Probabilidad $_{\text{calculada}}$ = .001 < Alfa = .05 / 2 = .025 por cola.

Valor $_{\text{calculado}}$ > Valor $_{\text{Critico}}$.

Por lo tanto, existe una diferencia estadísticamente significativa: se rechaza la hipótesis nula.

A nivel de las poblaciones:

Comparación de poblaciones: $\mu_{\text{niñas}} = \mu_{\text{niños}}$ (sin diferencia).

Por lo tanto, no existe una diferencia y la hipótesis nula es cierta, pero se rechazó (Esto es el Error Tipo I).

Volviendo al ejemplo anterior, la diferencia es que existiría el Error Tipo I cuando se compararan dos promedios y hay una diferencia estadísticamente significativa en las muestras cuando los promedios de las poblaciones son iguales. Si la nula es falsa, entonces es imposible cometer el Error Tipo I. Para ilustrar lo anterior, se presenta el siguiente ejemplo.

Comparación de muestras: $\bar{x}_{\text{niñas}} \neq \bar{x}_{\text{niños}}$.

Probabilidad $_{\text{calculada}}$ = .001 < Alfa = .05 / 2 = .025 por cola.

Valor $_{\text{calculado}}$ > Valor $_{\text{Critico}}$.

Por lo tanto, existe una diferencia estadísticamente significativa: se rechaza la hipótesis nula.

A nivel de las poblaciones:

Comparación de poblaciones: $\mu_{\text{niñas}} \neq \mu_{\text{niños}}$ (con diferencia).

Por lo tanto, existe una diferencia estadísticamente significativa en las muestras, y la hipótesis nula es falsa y se rechazó acertadamente: no hay error.

En resumen, existe una diferencia significativa entre los promedios de las muestras y provienen de poblaciones donde también había una diferencia entre sus promedios.

El segundo tipo de error es el *Error Tipo II* (β) que sucede cuando se falla en rechazar una hipótesis nula falsa. La probabilidad del Error Tipo II se denota con la beta: (β). Este Error Tipo II no es considerado un error como el Error Tipo I porque no se elige este primero (cf. Lane et al., 2014). De hecho, los Errores Tipo I y II tienen una relación inversa:

Si el α aumenta, el β disminuye y viceversa.

De nuevo el ejemplo de los promedios de niñas vs. niños, cuando se comparan dos promedios de dos muestras mediante un test estadístico, el resultado indica que no existe diferencia estadísticamente significativa, pero los promedios de las poblaciones de origen eran diferentes. Este ejemplo de las niñas vs.

niños es resumido enseguida (probabilidad $_{calculada}$ = .06 y un alfa .05 / 2 con dos colas) para ilustrar el Error Tipo II:

Comparación de muestras: $\bar{x}_{\text{niñas}} = \bar{x}_{\text{niños}}$

Probabilidad $_{calculada}$ = .06 > Alfa = .05 / 2 = .025 por cola.

Valor $_{calculado}$ < Valor $_{crítico}$

Por lo tanto, no existe una diferencia estadísticamente significativa: se falla en rechazar la hipótesis nula: i.e., la hipótesis nula era falsa y no se rechazó.

A nivel de las poblaciones:

Comparación de poblaciones: $\mu_{\text{niñas}} \neq \mu_{\text{niños}}$ (con diferencia).

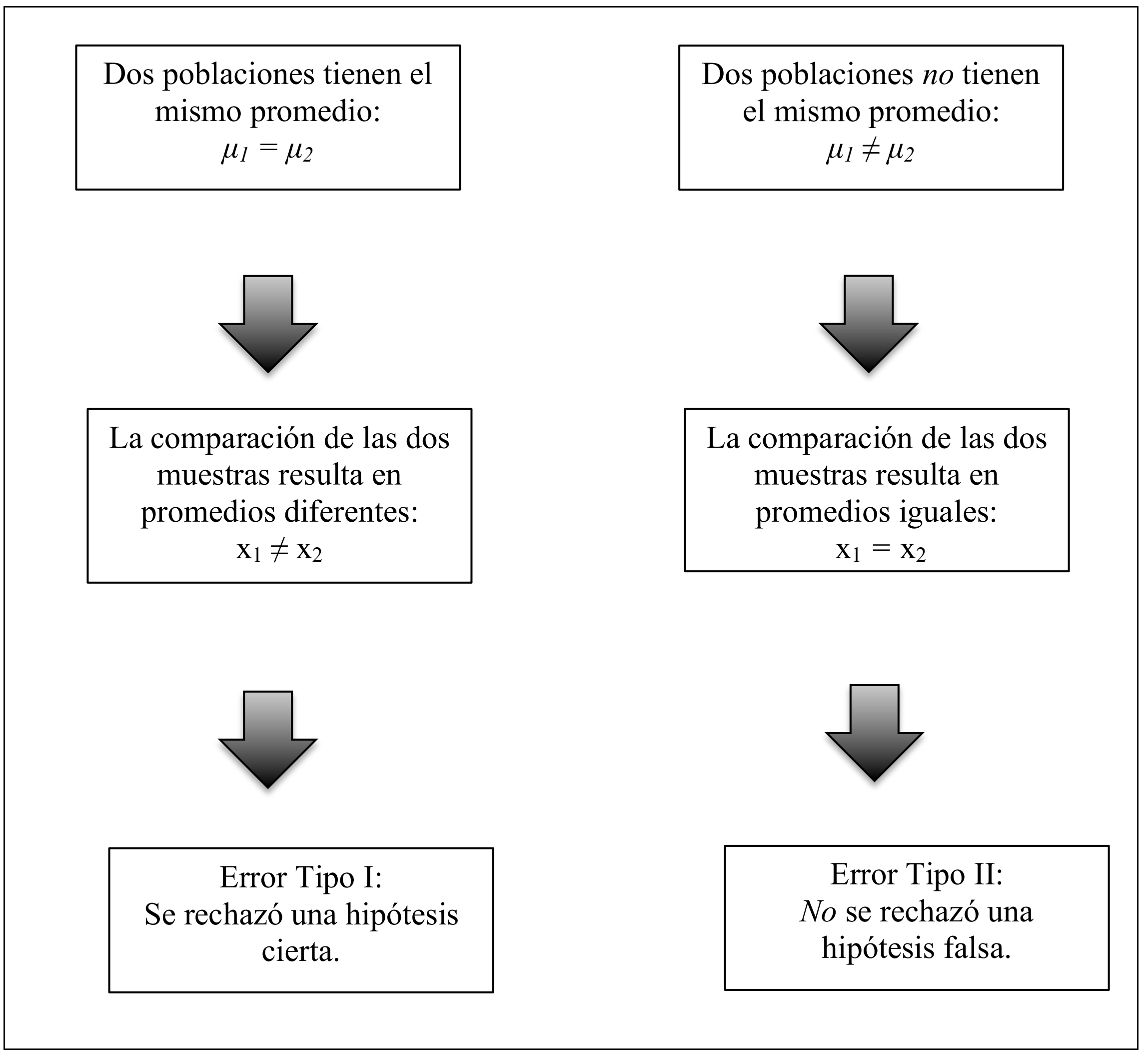
Por lo tanto, existe una diferencia entre las poblaciones de niños y niñas, pero no se rechazó la hipótesis nula (Esto es el Error Tipo II).

El *no* encontrar una diferencia estadísticamente significativa implica que los datos *no* proveen evidencia lo suficientemente fuerte de que la nula es falsa. En este caso, se concluye que los promedios de las muestras son iguales y se asume que los promedios de las poblaciones de origen también son iguales. Si resulta que los promedios de las poblaciones eran diferentes y se falló en rechazar la hipótesis nula, entonces se cometió el Error Tipo II.

La carencia de significancia estadística (probabilidad $_{calculada}$ > Alfa) no apoya la conclusión de que la nula es cierta (Lane et al., 2014). Por lo tanto, una investigadora o investigador no debe de cometer el error de concluir incorrectamente que la nula es cierta cuando el test no fue significativo. Lane et al. (2014) recomendó que una mejor opción es considerar el test *inconcluso*. En pocas palabras, el Error Tipo II solo puede ocurrir si la nula es falsa. De nuevo, si la nula es falsa, entonces la probabilidad del error Tipo II se llama beta (β). La probabilidad de correctamente rechazar la nula es $1 - \beta$ lo cual se *llama Poder Estadístico* que es la probabilidad de rechazar una nula cuando es falsa. El Poder Estadístico será explicado más adelante

La Tabla 8.1 muestra un resumen de los errores Tipo I y II. Para el Error Tipo I, se tiene que no hay diferencia entre los promedios de dos poblaciones (i.e., $\mu_1 = \mu_2$). Se toman las muestras de cada una de estas poblaciones y se comparan. Resulta que sus promedios son diferentes ($\bar{x}_1 \neq \bar{x}_2$). Al rechazar la hipótesis nula, se comete el Error Tipo I porque se rechazó una nula que era cierta ya que no había diferencia en los promedios a nivel de las poblaciones. Por otro lado, se comete el Error Tipo II cuando se tiene dos poblaciones que difieren en sus promedios ($\mu_1 \neq \mu_2$), y, al comparar sus respectivas muestras, estas tienen promedios iguales ($\bar{x}_1 = \bar{x}_2$). Entonces, no se rechazó una hipótesis nula falsa porque los promedios de las poblaciones eran de hecho diferentes.

Tabla 8.1 Resumen de los Errores Tipo I y II.



Gráficamente, el Error Tipo I es mostrado al tener dos curvas de dos poblaciones (i.e., la 1 y la 2) que se sobre imponen perfectamente la una en la otra cuando los promedios de sus respectivas muestras están separados (Figura 8.1). Es decir, los promedios de estas dos poblaciones son iguales. Por otro lado, cuando se obtiene una muestra de cada una de las poblaciones, las distribuciones de estas están casi en los extremos de cada población. Al estar las muestras tan separadas, sus promedios son estadísticamente y significativamente diferentes cuando una población se sobre empalma en la otra.

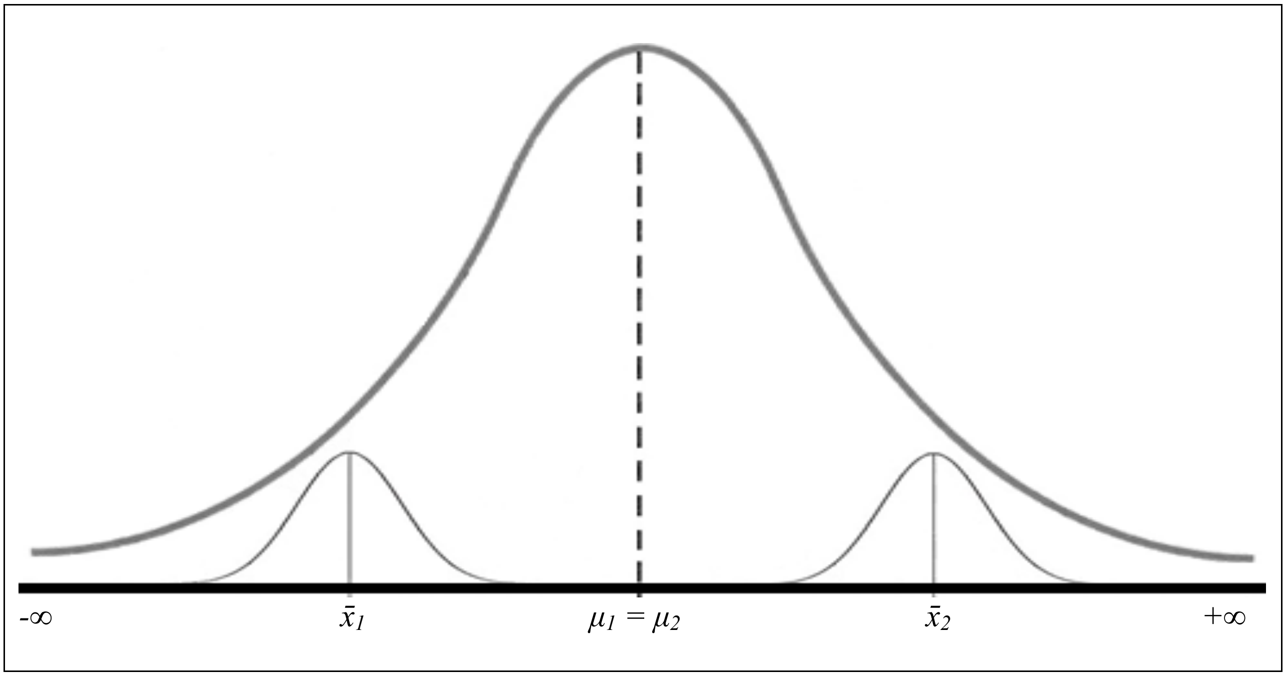


Figura 8.1 **Error Tipo I.**

El Error Tipo II es mostrado al tener dos curvas de dos poblaciones (i.e., la 1 y la 2) que están separadas (Figura 8.2) mientras los promedios de sus respectivas muestras se empalman. Es decir, los promedios de estas dos poblaciones no son iguales. Por otro lado, cuando se obtiene una muestra de cada una de las poblaciones, las distribuciones de estas están sobre puestas. Al estar las muestras empalmadas, sus promedios no son estadísticamente y significativamente diferentes.

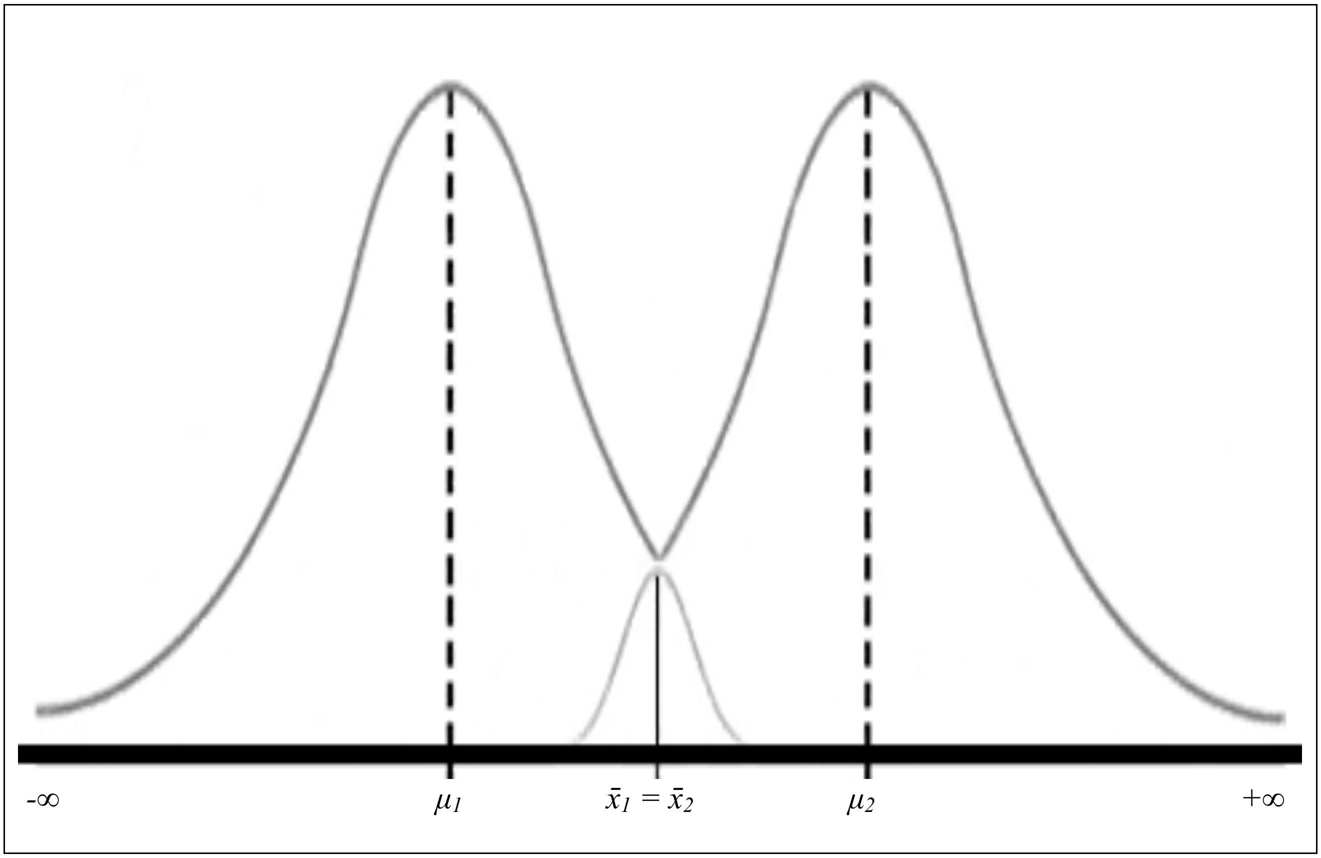


Figura 8.2 **Error Tipo II.**

Los Errores Tipo I y II tienen una relación inversa como se verá más adelante. Cuando la probabilidad de cometer el Error Tipo I incrementa, el Error Tipo II disminuye. En términos absolutos, no hay un error que sea más importante que el otro. Habría que ver el contexto en particular para decir si se decide caer en la probabilidad de cierto riesgo. También, estos errores pueden ser clasificados como *Falso Positivo* y *Falso Negativo* (Tabla 8.2).

Tabla 8.2 Falso Positivo vs. Falso Negativo.

Tipos de Error	Ejemplos Hipotéticos
Falso Positivo Error Tipo I Se rechaza una Hipótesis nula Cierta	Dos muestras de dos diferentes países (México vs. Guatemala) toman un examen de conocimientos como el de la prueba PISA (Programa Internacional para la Evaluación de Estudiantes). Los resultados del análisis de la comparación de los puntajes promedio de las muestras enseñan que hay una diferencia estadísticamente significativa entre ambos. Sin embargo, esta diferencia solo se debe a los resultados de las muestras porque si se compararán los promedios de las poblaciones estas serían iguales. En otras palabras, el Falso positivo seria el catalogar a la población de estudiantes como superiores en puntajes del PISA que los estudiantes mexicanos.
Falso Positivo	Declarar a un estudiante con el atributo de aptitudes sobresalientes mientras en realidad es un estudiante con atributos promedio.
Falso Negativo Error Tipo II No se rechaza una Hipótesis nula Falsa	Dos muestras de dos diferentes países (México vs. Estados Unidos de Norteamérica) toman un examen de conocimientos. Los resultados del análisis de la comparación de los puntajes promedio de las muestras enseñan que *no* hay una diferencia estadísticamente significativa entre ambos. Esta falta de diferencia se debe a los resultados de las muestras porque cuando se compararán los promedios de las poblaciones estas no serían iguales. En otras palabras, el Falso negativo seria el catalogar a la población de estudiantes como iguales en los puntajes del PISA cuando no lo son.
Falso Negativo	No declarar a un estudiante con el atributo de discapacidad intelectual cuando de acuerdo a sus capacidades si debía de haber sido declarado como tal.

Preguntas para Reflexionar

1.- ¿Cómo saber si se cae en el Error Tipo I y II?

2.- ¿Cómo se podría remediar esta situación de los Errores Tipo I y II?

3.- ¿Cuál de los dos errores es más importante de los dos?

4.- En qué se puede uno basar para determinar si un error es más importante que el otro?

Opinión del Autor

La enseñanza de esta sección es que se pueden cometer errores al generalizar los resultados de las muestras a las poblaciones. Hay que estar consciente de esta posibilidad y evitar un lenguaje que afirme que la hipótesis nula es falsa de acuerdo a la evidencia. En realidad, no se sabe si la hipótesis nula es falsa o cierta con un 100%. Esto es porque difícilmente se va a tener la información de algún parámetro de interés de una población.

Capítulo 9. Interpretando Resultados Significativos.

Los test o pruebas estadísticas se usan con las muestras y no con las poblaciones. Es decir, se comparan grupos o se correlacionan variables tomadas de muestras. Si estas muestras fueron tomadas al azar, cabe la posibilidad de poder hacer generalizaciones a las poblaciones de donde se tomaron. Unas de las interpretaciones *no* apropiadas para la significancia estadísticas son:

- creer que las hipótesis se pueden comprobar (una hipótesis se puede poner a prueba, pero nunca se puede confirmar porque difícilmente se tiene acceso a toda la población para hacerlo).
- generalizar a la población sin considerar la posibilidad de un error (Error Tipo I y II);
- sustentar que la probabilidad calculada es la magnitud del efecto práctico (para la magnitud del efecto se usa un tamaño del efecto como el *d* de Cohen explicado más adelante).

Volviendo al ejemplo de los promedios de calificaciones de los niños vs. las niñas, se tienen ambas poblaciones de origen y se quiere saber si hay evidencia de un fenómeno. En este caso, se podría inferir que el fenómeno es el efecto que tiene el sexo en las calificaciones. Cohen (1988, p. 2) explico que un investigador o investigador puede inferir la existencia de un fenómeno de dos maneras diferentes:

1.- Se asume que un fenómeno existe si el parámetro de una población y de otra difieren. Con el ejemplo anterior, el efecto del sexo en las calificaciones existiría si los promedios de las poblaciones de niños y niñas difirieran. En este caso, no se especifica si uno de los promedios es más grande que el otro. Por lo tanto, se están usando las dos colas de la distribución porque no se declara cuál de ellos va a ser más grande. Las hipótesis nula y alternativa quedarían de la siguiente forma:

H_0: $\mu_{\text{niñas}} = \mu_{\text{niños}}$

H_A: $\mu_{\text{niñas}} \neq \mu_{\text{niños}}$

La Figura 9.1 muestra una curva normal que representa la *diferencia* entre dos promedios de las poblaciones (i.e., $\mu_{\text{población 1}} - \mu_{\text{población 2}}$). Por el momento, solo es necesario explicar que esta curva va desde cero cuando la diferencia entre dos promedios no existe (i.e., $\mu_{\text{población 1}} - \mu_{\text{población 2}} = 0$) hasta el infinito negativo en cierta dirección y al infinito positivo en la dirección opuesta. Si se selecciona un alfa de .05 cuando se tienen dos colas, este alfa se divide entre dos: .05 / 2 = .025. Lo mismo parecía con otros coeficientes alfa que tendrían que ser divididos entre dos para encontrar el valor critico en alguna tabla de calores críticos (e.g., de valores *z*; valores *t*, etc.).

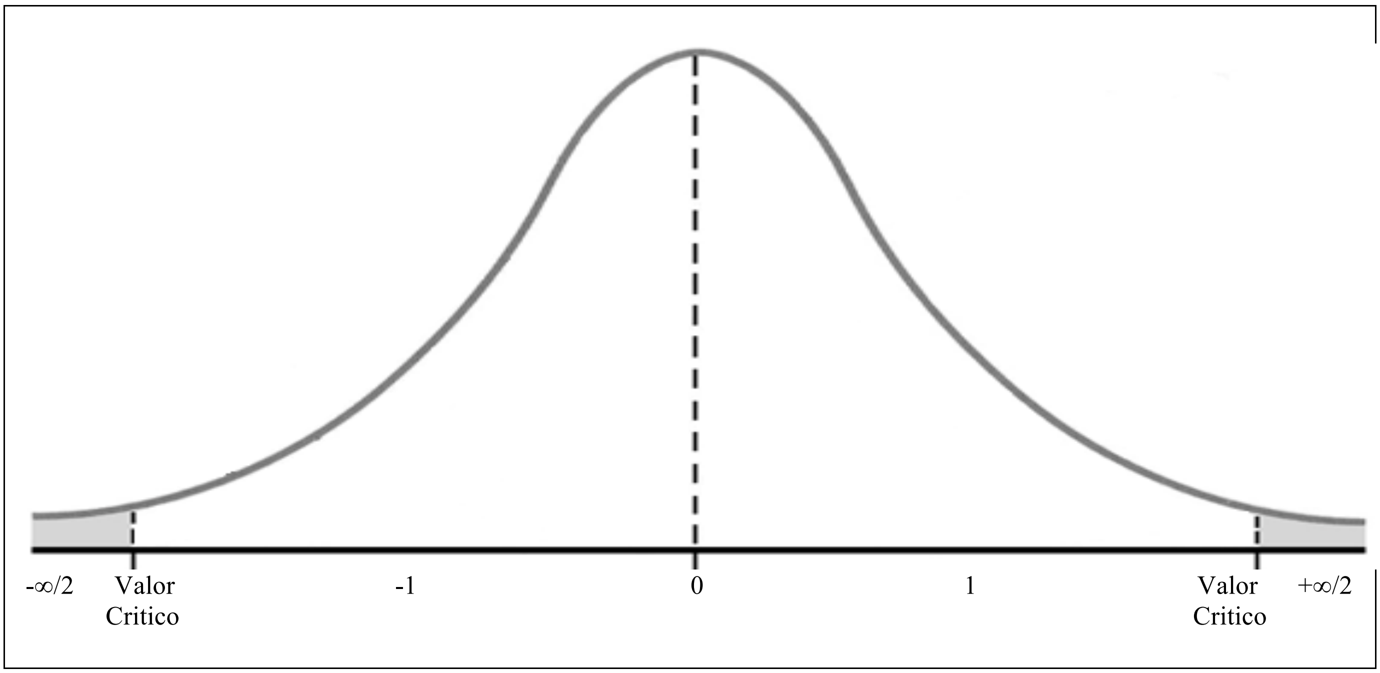


***Figura 9.1* Curva de la Diferencia entre Promedios de dos Poblaciones con dos Colas.**

2.- Se asume que un fenómeno existe si el parámetro como el promedio de una población y de otra difieren en cierta dirección previamente especificada. Un ejemplo de lo anterior seria que el promedio de la población de las niñas es mayor que el de los niños. En este caso, solo se está usando una de las colas de la distribución y en la medida que los promedios difieran en la dirección establecida se considera como evidencia de la diferencia. En hipótesis, lo anterior se expresaría así:

H_0: $\mu_{\text{niñas}} = \mu_{\text{niños}}$

H_A: $\mu_{\text{niñas}} > \mu_{\text{niños}}$

La Figura 9.2 muestra una curva normal que representa la diferencia entre dos promedios de las poblaciones (i.e., $\mu_{\text{población 1}} - \mu_{\text{población 2}}$). En esta curva, solamente se utiliza una cola porque se establecería en la hipótesis alternativa que un promedio es mayor que el otro:

i.e., H_A: $\mu_{\text{población 1}} > \mu_{\text{población 2}}$.

Esta curva va desde cero cuando la diferencia entre dos promedios no existe hasta el infinito positivo:

i.e., $\mu_{\text{población 1}} - \mu_{\text{población 2}} = 0$.

Si se selecciona un alfa de .05, este se utiliza en una cola para encontrar el valor critico en alguna tabla de valores críticos (e.g., de valores *z*; valores *t*, etc.).

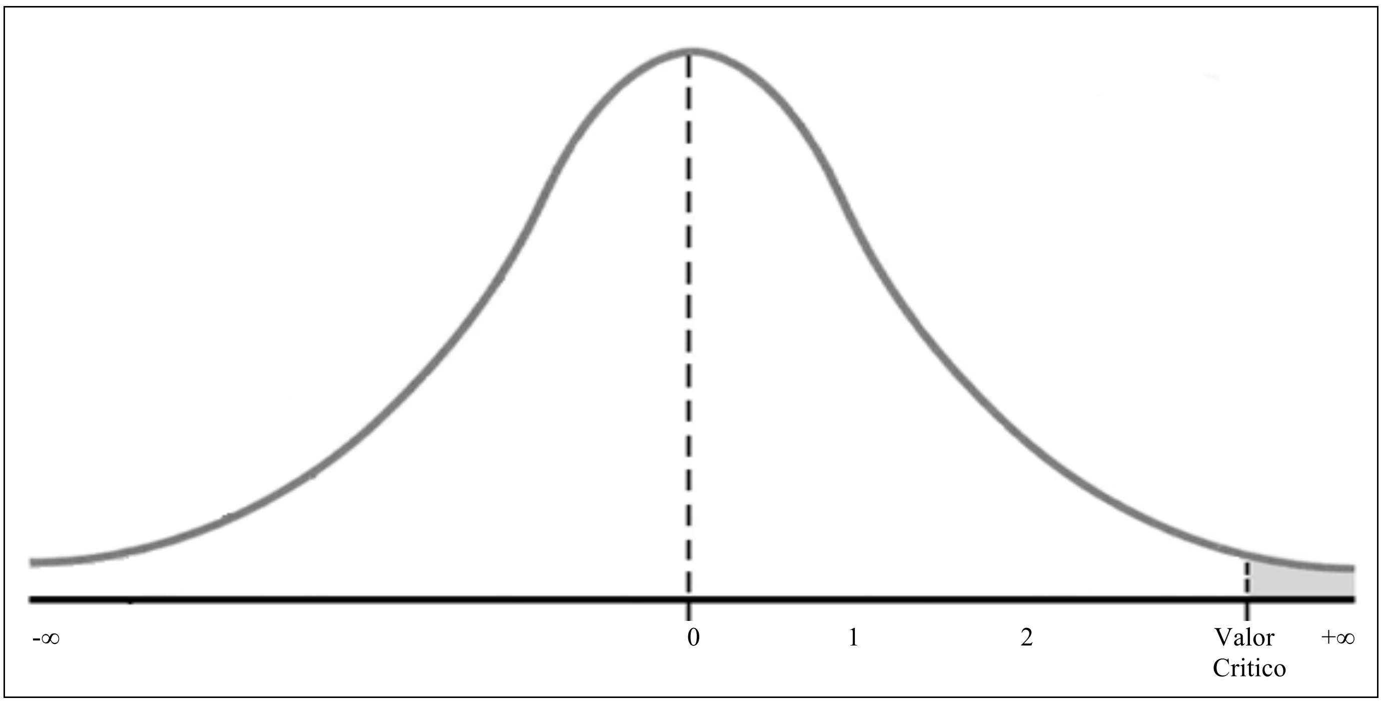


Figura 9.2 **Curva de la Diferencia entre Promedios de dos Poblaciones con una Cola.**

Otro ejemplo de la dirección de la diferencia puede ser uno en medicina. Se tienen dos grupos en un diseño experimental en el cual el grupo control recibe un placebo y el grupo de tratamiento recibe un nuevo medicamento. El medicamento es una nueva pastilla para aliviar el dolor de cabeza. Antes de empezar el experimento, los dos grupos son equivalentes porque en promedio les duele la cabeza al mismo nivel. Se esperaría que al cabo de finalizar el experimento el promedio entre ambos grupos fuera diferente porque unos recibieron el medicamento y otros no. Es decir, se esperaría que en promedio el grupo tratamiento sufriera menos intensidad de dolores de cabeza que el tratamiento. Entonces, esta diferencia sería una evidencia de que existe el fenómeno. El fenómeno es el efecto de la pastilla en los dolores de cabeza.

Otro aspecto para definir es el nivel de significancia. Según Cohen (1988, p. 3), un alfa igual a .05 significa dos cosas:

La primera: El fenómeno que se está cuestionando su existencia se manifiesta con una diferencia entre los valores de algún parámetro. De nueva cuenta este fenómeno puede ser el efecto del sexo en las calificaciones de matemáticas de las poblaciones de niñas y niños. En pocas palabras lo anterior seria:

$\mu_{\text{niñas}} - \mu_{\text{niños}} \neq 0$

La segunda: El estándar de prueba es un resultado de una muestra que ocurriría menos del 5% de las veces si la hipótesis nula es cierta. Probablemente, el ejemplo antes mencionado del experimento con la medicina con la pastilla del dolor de cabeza sea más apropiado. Si la hipótesis nula es cierta, quiere decir que la población de la muestra control y la población de tratamiento son iguales en su promedio de intensidad de dolores de cabeza:

$H_0: \mu_{\text{control}} - \mu_{\text{tratamiento}} = 0$

Suponiendo, el resultado del experimento da que el dolor de cabeza del grupo tratamiento es menor en promedio que el grupo control, y esta diferencia exhibe una probabilidad calculada de .01 que es menor al alfa seleccionado del .05. En pocas palabras:

Resultado: probabilidad $_{\text{calculada}}$ (.01) < alfa (.05).

Entonces, $\bar{x}_{\text{tratamiento}} < \bar{x}_{\text{control}}$

Esta diferencia ente los promedios de las muestras es evidencia de que existe el fenómeno: efecto del medicamento en el dolor de cabeza. Esta conclusión de que el efecto existe tiene un riesgo que es el Error Tipo I que es una probabilidad no mayor al 5%.

Cuando poner a prueba una hipótesis nula, Cohen (1988, p. 4) explico que es necesario tomar en cuenta el *Poder Estadístico de un Test para la Hipótesis Nula* (*the power of a statistical test of a null hypothesis*). Para simplicidad, se le llamara a este concepto: *poder estadístico*. Según Cohen (1988, p. 4), el poder estadístico

"…es la probabilidad de un test que llevara correctamente al rechazo de una hipótesis nula, i.e., la probabilidad de que el test resultara en la conclusión de que el fenómeno existe." Asimismo, Hinkle et al. (2003, p. 738) definió elegantemente que el poder estadístico es la probabilidad de rechazar la hipótesis nula cuando esta es falsa. Por ejemplo, el escenario de una hipótesis nula es *falsa* es cuando los promedios de las poblaciones son diferentes, pero los promedios de las muestras son iguales:

Esto es lo que sucede a nivel de los promedios de las poblaciones: $\mu_1 \neq \mu_2$.

Sin embargo, en la muestra se encuentra que: H_0: $\bar{x}_1 = \bar{x}_2$.

El poder estadístico depende de tres elementos: el alfa que se seleccione, el tamaño del efecto, y el tamaño de la muestra. En otras palabras, el poder estadístico está en función de estas tres variables (Cohen, 1988, p. 14):

Poder Estadístico (α, d, n).

El Poder Estadístico está en función del alfa, del d y del tamaño de la muestra.

α = alfa seleccionado (e.g., .01, .05, o .10).

d = este es un ejemplo de un efecto: d de Cohen.

n = tamaño de la muestra.

Cohen (1988) también explico que cada una de estas variables está en función del resto. Esto quiere decir que si tres de las variables están fijas (i.e., la barra encima de variable indica que esta fija: no varía), la cuarta restante está determinada por estas primeras:

Poder Estadístico ($\bar{\alpha}$, $\bar{d}$, $\overline{n}$) .

$\alpha = (\bar{n}, \bar{d}, \overline{\text{Poder Estadístico}})$.

$d = (\bar{\alpha}, \bar{n}, \overline{\text{Poder Estadístico}})$.

$n = (\bar{\alpha}, \bar{d}, \overline{\text{Poder Estadístico}})$.

También, el poder estadístico tiene una relación inversa con la probabilidad de cometer el Error Tipo II (β): Cuando el Error Tipo II aumenta, el poder estadístico disminuye. La fórmula es que el poder estadístico

= 1 - β. En otras palabras, el área de la probabilidad del poder estadístico es lo que queda una vez que se substrae el Error Tipo II.

Para tener una referencia del nivel del poder estadístico, Cohen (1988) recomendó que se tenga un poder estadístico de por lo menos un 80% de probabilidad de rechazar la hipótesis nula cuando sea falsa. Esto implica que el Error Tipo II sea del 20%. Para una serie de ejemplos heurísticos de la relación del poder estadístico, el alfa, el tamaño de la muestra, el *d*, y el coeficiente beta, se usó el software de *Gpower* (http://www.gpower.hhu.de/), los cuales se muestran en las próximas secciones: Análisis con Dos Cola y Análisis con una Cola. Este software permite calcular a priori el tamaño de una muestra para un estudio y también puede ser usado después del análisis de un estudio para estimar el poder estadístico. También, permite calcular el poder estadístico *post hoc*.

Preguntas para Reflexionar

1.- ¿Qué femémonos educativo se cree comúnmente que se manifiestan en las poblaciones?

2.- ¿Qué se haría cuando se encuentra una probabilidad calculada de .049 cuando se adoptó un alfa de .05?

3.- ¿En qué ocasiones de investigación educativa se usa un test de una cola en lugar de dos colas?

4.- ¿Cuál de las variables relacionas al poder estadístico tiene más peso en cambiarlo?

Opinión del Autor

Además de rechazar o no una hipótesis nula cuando los análisis indican, esta sección empieza a explicar cómo las hipótesis alternativas pueden indicar el sentido en el que dos promedios pueden ser diferentes: e.g., el promedio uno > promedio dos. Esto es importante en la investigación educativa que usa diseños experimentales, en los cuales se asume que habrá una mejora en algún aspecto después de algún tratamiento educativo. Un diseño experimental sería uno en el que un grupo de estudiantes sea evaluado (pre-test) para ver dónde están sus conocimientos de matemáticas. Luego, se les daría un tratamiento con el propósito de mejorar sus conocimientos matemáticos. Al último, se evaluarían sus conocimientos (post-test) de nueva cuenta para comparar el pre y el post-test. En este caso, se esperaría que el promedio del grupo en el post-test sería más alto que el previo. No tendría mucho sentido que se tuviera una hipótesis alternativa donde no hubiera dirección de cual promedio iba a ser más grande.

Una situación más es: la hipótesis nula es cierta cuando no se rechaza. El crearla cierta implicaría aceptarla. Pero no se dice que se acepta la nula cuando no se rechaza. Entonces, surge la pregunta: ¿Cómo deben de interpretarse los resultados *no* significativos? De este modo se interpretan: Se reporta que no hay suficiente evidencia de diferencia entre los promedios. Unos autores dicen que el test se debe de declarar inconcluso. También, está la situación que se tiene un test con muestras y que no haya diferencia entre los promedios de estas, no quiere decir necesariamente que no exista diferencia entre los promedios de las poblaciones.

Capítulo 10. Relación de un Test de Dos Colas y Una Cola con Otras Variables.

Un o una investigadora decide si establece su hipótesis alternativa, afirmando que dos promedios de dos muestras son diferentes (i.e., dos colas) o, por otro lado, que un promedio es mayor que el otro (una cola). La primera posibilidad es que la hipótesis alternativa solo establezca que hay una diferencia entre el promedio uno y el dos, sin decir cuál será mayor. Para este tipo de test, se usan las dos colas de una distribución porque no se estableció el sentido de la hipótesis alternativa. Posiblemente, sea más adecuado usar un test de dos colas cuando se realiza algún tipo de análisis exploratorio donde no se sabe mucho del efecto de algo. Por ejemplo, si se está estudiando el efecto en el aprendizaje del uso del teléfono durante una clase en estudiantes de una universidad, no se sabe si el uso de este tendrá un efecto en detrimento o a favor del aprendizaje. Lo anterior es asumiendo que no hay muchos precedentes que indiquen algún efecto. Entonces, se usaría una hipótesis alternativa, en la cual no se establece un sentido: i.e., el promedio de los que usan el teléfono en clase es diferente a los que no usan. Por otro lado, si se asume que hay precedentes para el ejemplo anterior y que los usuarios de teléfonos en clase aprenden menos que los no usuarios, se puede establecer la hipótesis alternativa como: el promedio de los que usan el teléfono en clase es menor al resto (i.e., una sola cola).

a. Análisis con Dos Colas.

La Tabla 10.1 muestra doce escenarios, en los cuales se usa un análisis con dos colas fijas, así como el tamaño de la muestra esta fija. Para esta serie de ejemplos se compararon los promedios de dos muestras con el mismo número de participantes 25 y 25 (n = 50) con el test t de estudiante de muestras independientes. Para más información acerca de este test, se recomienda a Hinkle et al. (2003). Los doce escenarios están divididos en *tres* grupos de cuatro de acuerdo con un coeficiente alfa: .01; .05; y .10. Asimismo, el d de Cohen va de un valor de 0.20 a 1.00 para cada uno de estos tres grupos. El d de Cohen es un efecto que mide la magnitud de la diferencia entre dos promedios (este efecto será explicado más adelante en el Capítulo 11). Al incrementar el alfa y el *d*, el coeficiente beta y el poder estadístico cambian (esto se ilustra más adelante). Se parte de la hipótesis nula que no hay diferencia entre los promedios de dos muestras. Las hipótesis son las siguientes:

H_0: $\bar{x}_1 = \bar{x}_2$

Con dos colas, la hipótesis alternativa no tiene dirección: H_A: $\bar{x}_1 \neq \bar{x}_2$

Los coeficientes alfa: .01; .05; y .10.

$n = 50$

En la Tabla 10.1, los resultados del *Gpower* muestran que al incrementar secuencialmente el valor del coeficiente alfa de .01 a .05 y .10, y manteniendo el *d* fijo (e.g., $d = 0.20$; escenarios: I., V., y IX.), el coeficiente beta progresivamente disminuye (i.e., .95, .84, y .74) y, por lo tanto, el poder estadístico aumenta (i.e., .05, .16, y .26):

Manteniendo el *d* fijo, si el α aumenta, el β disminuye y el poder estadístico aumenta.

Recordando: Poder estadístico = $(1 - \beta)$.

De una manera similar, cuando el *d* de Cohen aumenta, manteniendo fijo el alfa, el coeficiente beta disminuye y el poder estadístico aumenta. Uno ejemplo de lo anterior es cuando, en los escenarios del I. Al IV. (Tabla 10.1) con un alfa fijo al .01, el *d* va aumentando (i.e., 0.20, 0.50, .080, y 1.00), el coeficiente beta disminuyendo (.95, .55, .09, y .01), y el poder estadístico incrementando (.05, .45, .91, y .99). En pocas palabras, lo siguiente se puede resumir:

Manteniendo el α, si el *d* aumenta, el β disminuye y el poder estadístico aumenta.

Esto implica que un aumento del *α y d* está asociado con una disminución en el Error Tipo II (fallar en rechazar una H_0 cuando es falsa), céteris páribus. El coeficiente alfa puede ser manipulado por la investigadora o investigador. Sin embargo, posiblemente tenga que justificar el uso de un alfa que no sea el tradicional: .05 a la hora de publicar o presentar un manuscrito. Por otro lado, el *d* es una variable que un investigador o investigadora no pueden manipular, sino que resulta del análisis de los datos.

Tabla 10.1 Dos Colas, *d* de Cohen, *n*, alfa, beta y Poder Estadístico.

Escenario	Colas	*alfa* (α)	*d*	*n*	*beta* (β)	Poder Estadístico (1 - *beta*)
I.	2	.01	0.20	50	.95	.05
II.	2	.01	0.50	50	.55	.45
III.	2	.01	0.80	50	.09	.91
IV.	2	.01	1.00	50	.01	.99
V.	2	.05	0.20	50	.84	.16
VI.	2	.05	0.50	50	.31	.69
VII.	2	.05	0.80	50	.03	.97
VIII.	2	.05	1.00	50	.01	.99
IX.	2	.10	0.20	50	.74	.26
X.	2	.10	0.50	50	.21	.79
XI.	2	.10	0.80	50	.01	.99
XII.	2	.10	1.00	50	.001	.99

Las Figuras 10.1, 10.2, y 10.3 (generadas en *Gpower*) muestran gráficamente la relación del α y *d* con el β y con el poder estadístico que se presenta en la Tabla 10.1 (i.e., mismos 12 escenarios y mismo análisis de comparación de promedios de dos muestras independientes). En los Escenarios I. al XII., se puede observar que las dos muestras que se están comparando poseen distribuciones normales. Para la simplicidad de los escenarios, se puede decir que la muestra 1 está a la izquierda y la muestra 2 a la derecha. En el escenario I., el *d* es la distancia que hay entre el promedio de la muestra uno a la muestra dos. Esta distancia es de 1 / 5 (0.20) de desviación estándar. El coeficiente alfa fijo (.01) es la parte sombreada que aparece en las colas de la muestra 1: i.e., .01 / 2 = .005 por cada cola. Por otro lado, la parte sombreada de la muestra dos es el coeficiente beta con un área de .95 o en porcentaje: 95% de la probabilidad de cometer el Error Tipo II. De hecho, esta área sombreada exhibe la letra griega: β. De nuevo, el poder estadístico se calcula con la diferencia entre 1 (i.e., 100% de la probabilidad) y el coeficiente β: 1 - .95 = .05 de poder estadístico. En resumen, con el Escenario I se tendría más probabilidad de cometer el Error Tipo II que es fallar en rechazar la hipótesis nula cuando es falsa en la población que en los otros 11 escenarios donde se tienen menos probabilidad del Error Tipo II, y, por lo tanto, más probabilidad del poder estadístico. La razón es que el escenario I. tiene la combinación del alfa y el *d* más pequeños de todos los escenarios. Esta situación del Escenario I se puede resumir de la siguiente manera:

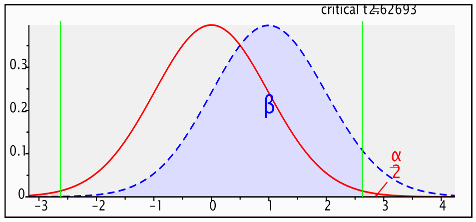
H_0: $\bar{x}_1 = \bar{x}_2$

Dado un α = .01 / 2 por cola y un *d* = 0.20, se tiene una probabilidad de 95% de fallar en rechazar una hipótesis nula falsa cuando se asume lo siguiente: $\mu_1 \neq \mu_2$

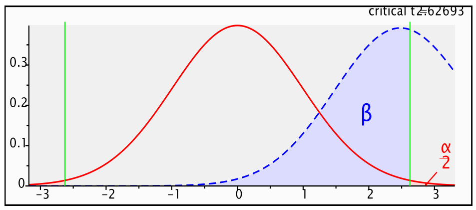
Conforme se va del escenario I. al IV., la distancia entre la muestra 1 y la 2 va incrementando dado el *d*, área sombreada del coeficiente beta va disminuyendo (Figura 10.1). Es decir, la probabilidad de cometer el Error Tipo II disminuye y el poder estadístico aumenta. Desde el Escenario III, el poder estadístico (97%) ha sobrepasado el umbral que estableció Cohen (1988) de 80%. De la misma manera, el *d* aumenta en las Figura10.2 y 10.3 con un alfa = .05 y a .10, y el coeficiente beta disminuye y el poder estadístico aumenta.

Las Figuras 10.2 y 10.3 contienen escenarios similares a los ya descritos en la Figura 10.1 para representar la Tabla 10.1. La diferencia estriba en que el alfa es diferente ahora (.05 para la Figura 10.2 y .10 para la Figura 10.3). Dada esta diferencia, el incremento en el alfa en combinación con el cambio en el *d* hacen que el coeficiente beta disminuya más rápido que en la Figura 10.1.

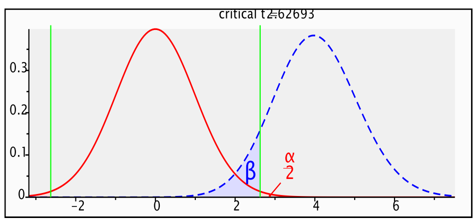
Escenario I



Escenario II



Escenario III



Escenario IV

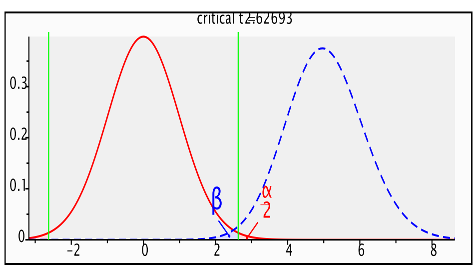
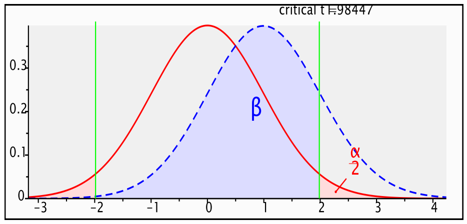
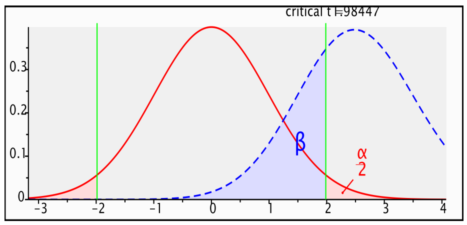


Figura 10.1 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .01 para dos Colas.**

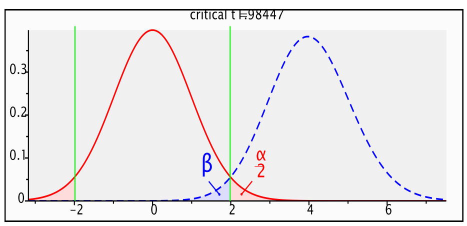
Escenario V



Escenario VI



Escenario VII



Escenario VIII

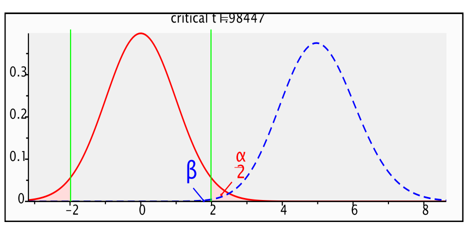
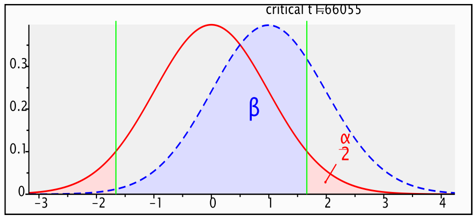
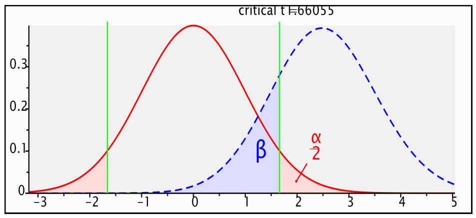


Figura 10.2 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .05 para dos Colas.**

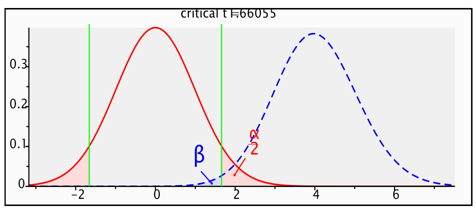
Escenario IX



Escenario X



Escenario XI



Escenario XII

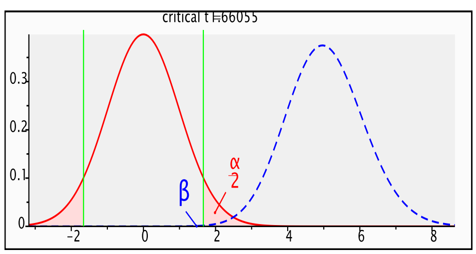
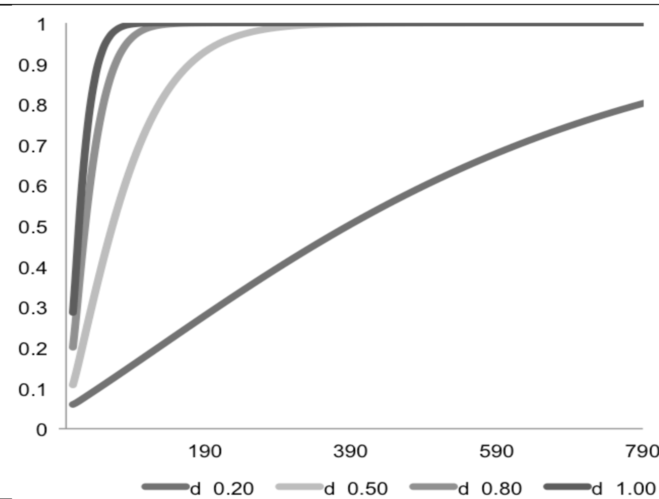


Figura 10.3 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .10 para dos Colas.**

b. Análisis con Dos Colas y Tamaño de la Muestra.

Para la planeación de un estudio, el software *Gpower* sirve para calcular el tamaño de una muestra para obtener ese 80% que recomendó Cohen (1988). Esta planeación es posible para estudios donde se comparen promedios (e.g., test *t* de estudiante, análisis de la varianza, etc.) o se relacionen variables (e.g., análisis de regresión). Por ejemplo, el poder estadístico puede ser calculado de acuerdo al alfa que se seleccione, el tamaño de la muestra y el *d* de Cohen que podría ocurrir. Más al respecto, la Figura 10.4 muestra la relación de la muestra y el poder estadístico dados cinco niveles de *d* (i.e., 0.20, 0.50, 0.80, y 1.00) y un alfa = .05. Este análisis también es para comparar promedios con un test *t* de estudiante de muestras independientes del mismo tamaño (e.g., muestra uno con 25 participantes y muestra 2 con 25 participantes). El eje de la *x* es para el tamaño de la muestra (i.e., de 0 a 790) y el eje de *y* es para el poder estadístico (de 0 a 1). Es decir, el poder estadístico está en función de la muestra, manteniendo fijos el α y el *d*: recordando, Poder estadístico $(n, \bar{\alpha}, \bar{d})$.

Continuando con la Figura 10.4 y ejemplificando, si se estima que el valor del *d* será de 0.20, sería necesario considerar una muestra de 790 participantes, observaciones u objetos para alcanzar por lo menos un poder estadístico del 80%. En contraparte, si se estima que se tendrá un *d* = 1.00, solo se necesitan 40 participantes, observaciones u objetos para por lo menos alcanzar un 80% de poder estadístico.



d de Cohen	*n*	Poder Estadistico
0.20	790	.80
0.50	130	.81
0.80	60	.86
1.00	40	.87

Nota: El coeficiente alfa = .05. La hipótesis alternativa solo implica una diferencia entre los promedios de dos muestras: $\bar{x}_1 \neq \bar{x}_2$, i.e., usando las dos colas para los análisis. Las curvas señalan la relación positiva entre el tamaño de la muestra (*VI*) y el poder estadístico (*VD*).

Figura 10.4 **Planeación de la Muestra con el d de Cohen y el Poder Estadístico con Dos Colas.**

c. Análisis con una Cola.

La Tabla 10.2 muestra también doce escenarios en los cuales se usa un análisis con una cola fija, así como el tamaño de la muestra esta fijo (n = 50). También, se usa el test t de estudiante de comparación de dos muestras independientes. Las hipótesis de este ejemplo heurístico son:

H_0: $\bar{x}_1 = \bar{x}_2$

H_A: $\bar{x}_1 < \bar{x}_2$: Esta hipótesis alternativa implica una sola cola porque el promedio de la muestra 2 se espera que esta última sea mayor a la muestra 1.

Los coeficientes alfa son: .01; .05; y .10 que aparecen en tres diferentes grupos con cuatro escenarios por grupo como en el ejemplo anterior de las dos colas. Asimismo, el d de Cohen va de un valor de 0.20 a 1 por cada uno de los tres grupos. Ya sea que se mantenga fijo al α o al d cuando uno de los dos incrementa, y esto causa que el coeficiente beta disminuya, y, por lo tanto, el poder estadístico aumente. Las siguientes, dos conclusiones muestran lo anterior:

Manteniendo el d fijo, si el α aumenta, el β disminuye y el poder estadístico.

Nota: Ver los escenarios: XIII, XVII, y XXI como ejemplo para apreciar el cambio en el α, la disminución en el β y el aumento en el poder estadístico (Tabla 10.2).

Manteniendo el α, si el d aumenta, el β disminuye y el poder estadístico.

Nota: Ver los escenarios: XIII, XIV, XV y XVI como ejemplo para apreciar el cambio en α, la disminución en β y el aumento en el poder estadístico (Tabla 10.2).

En conclusión, un aumento del α está asociado con una disminución en el Error Tipo II (fallar en rechazar una H_0 cuando es falsa), céteris páribus. Además, el aumento del d tiene el mismo efecto que el α en la disminución del β y el aumento del poder estadístico. Por otro lado, el d es una variable que un investigador o investigadora no pueden manipular, sino que resulta del análisis de los datos, pero si puede aumentar el *alfa* aunque esto implica incrementar la probabilidad de cometer el Error Tipo I. También, puede elegir un análisis de una cola en lugar de dos, y cuando se elige una cola, se tendrá más poder estadístico (esta comparación se efectuará más adelante), céteris páribus.

Tabla 10.2 Una Cola, *d* de Cohen, *n*, alfa, beta y Poder Estadístico.

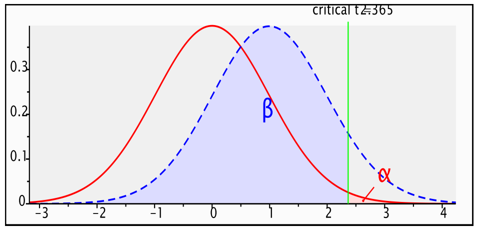
Escenario	Colas	*alfa* (α)	*d*	*n*	*beta* (β)	Poder Estadístico (1 - *beta*)
XIII.	1	.01	0.20	50	0.91	.09
XIV.	1	.01	0.50	50	0.45	.55
XV.	1	.01	0.80	50	0.06	.94
XVI.	1	.01	1.00	50	0.01	.99
XVII.	1	.05	0.20	50	.75	.25
XVIII.	1	.05	0.50	50	.21	.79
XIX.	1	.05	0.80	50	.01	.99
XX.	1	.05	1.00	50	.01	.99
XXI.	1	.10	0.20	50	.62	.38
XXII.	1	.10	0.50	50	.12	.88
XXIII.	1	.10	0.80	50	.01	.99
XXIV.	1	.10	1.00	50	.001	.99

Las Figuras 10.5, 10.6, y 10.7 muestran gráficamente la relación del α y *d* con el β y con el poder estadístico que se presenta en la Tabla 10.2 (i.e., mismos 12 escenarios y mismo análisis de comparación de promedios de dos muestras con un test *t* de estudiante de grupos independientes). En los escenarios XIII. al XXIV., se puede observar que las dos muestras que se están comparando poseen distribuciones normales. Para la simplicidad de los escenarios, se puede decir que la muestra 1 está a la izquierda y la muestra 2 a la derecha. En el escenario XIII., el *d* es la distancia que hay entre el promedio de la muestra uno la muestra dos. Esta distancia es de 1 / 5 (0.20) de desviación estándar. El coeficiente alfa fijo (.01) es la parte sombreada que aparece en las colas de la muestra 1. Por otro lado, la parte sombreada de la muestra dos es el coeficiente beta con un área de .91 o en porcentaje: 91% de la probabilidad de cometer el Error Tipo II. El poder estadístico se calcula con la diferencia entre 1 (i.e., 100% de la probabilidad) y coeficiente β: 1 - .91 = .09. En otras palabras, en el escenario XIII se tendría más probabilidad de cometer el Error Tipo II que es rechazar la hipótesis nula cuando es falsa que en los otros 11 escenarios. La razón es que el escenario XIII. tiene la combinación del alfa y el *d* más pequeños de todos los escenarios. Si se encuentra lo siguiente:

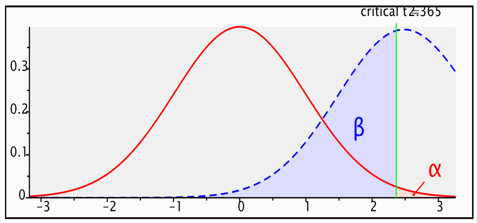
H_0: $\bar{x}_1 = \bar{x}_2$

Se tiene una probabilidad de 91% de fallar en rechazar una hipótesis nula falsa: $\mu_1 \neq \mu_2$. Conforme se va del escenario XIII. al XVI., la distancia entre la muestra 1 y la 2 va incrementando dado el *d*, área sombreada del coeficiente beta va disminuyendo (Figura 10.5). Es decir, la probabilidad de cometer el Error Tipo II disminuye y el poder estadístico aumenta. Desde el Escenario XV para el $\alpha = .01$, el poder estadístico (94%) ha sobrepasado el umbral que estableció Cohen (1988) de 80%. De la misma manera, que el *d* aumenta en la Figura 10.5 con un alfa = .01, el coeficiente beta disminuye y el poder estadístico aumenta. Lo mismo pasa en las Figuras 10.6 y 10.7. cuando el *d* incrementa con sus respectivos coeficientes alfas de .05 y .10 respectivamente y el coeficiente beta disminuye ocasionando que el poder estadístico aumenta.

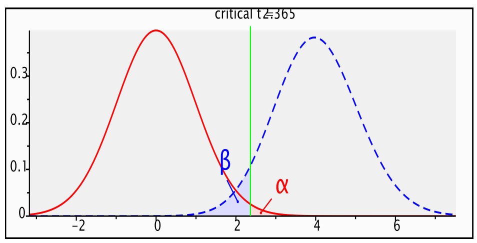
Escenario XIII



Escenario XIV



Escenario XV



Escenario XVI

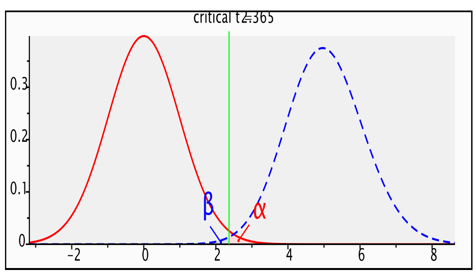
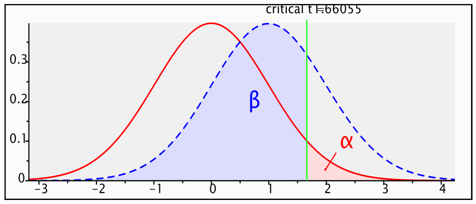
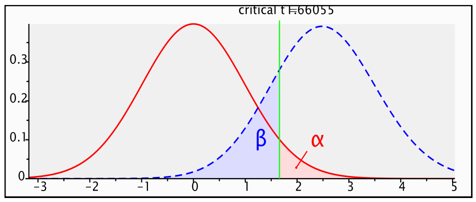


Figura 10.5 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .01 para una cola.**

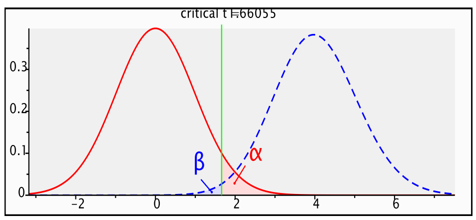
Escenario XVII



Escenario XVIII



Escenario XIX



Escenario XX

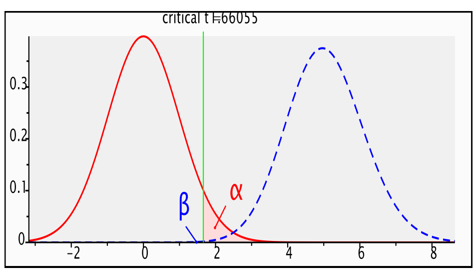
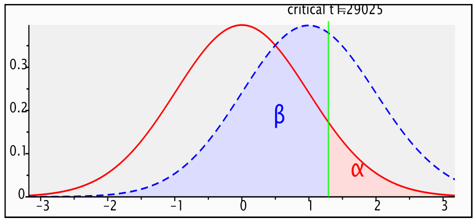
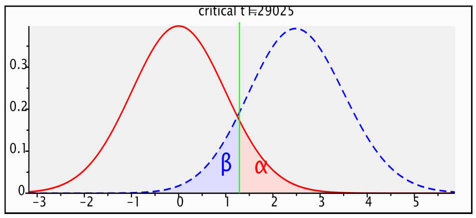


Figura 10.6 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .05 para una cola.**

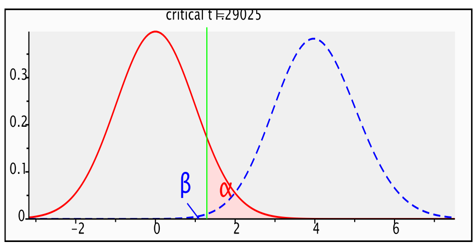
Escenario XXI



Escenario XXII



Escenario XXIII



Escenario XXIV

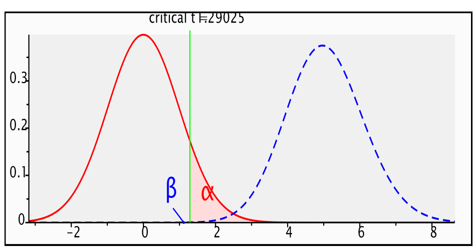
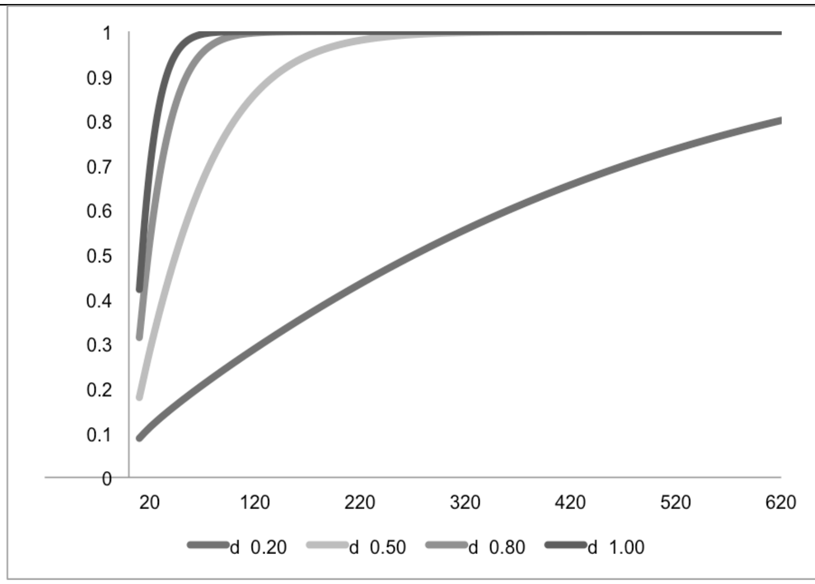


Figura 10.7 **Distancia entre la Muestra I y II, *d* de Cohen, Área del Coeficiente Beta y Poder Estadístico con un Alfa fijo = .10 para una cola.**

d. Una Cola y Tamaño de la Muestra.

Para la planeación de un estudio, el software *Gpower* sirve para planear el tamaño de una muestra. Más al respecto, la Figura 10.8 muestra la relación de la muestra y el poder estadístico dados cinco niveles de *d* (i.e., 0.20, 0.50, 0.80, y 1.00) y un coeficiente alfa fijo = .05. Este análisis también es para comparar promedios con un test *t* de estudiante de muestras independientes con un test de una sola cola. Por ejemplo, si se estima que el valor del *d* será 0.20, sería necesario considerar una muestra de 620 participantes, observaciones u objetos para llegar al nivel del 80% de poder estadístico. En contraparte, si se estima que se tendrá un *d* de 1.00, solo se necesitan 30 participantes, observaciones u objetos para llegar al 80% de poder estadístico.



d de Cohen	N	Poder Estadistico
0.20	620	.80
0.50	105	.82
0.80	45	.84
1.00	30	.85

Nota: El coeficiente alfa = .05. La hipótesis alternativa implica una diferencia en cierto sentido entre los promedios de dos muestras: $\bar{x}_1 < \bar{x}_2$. Las curvas señalan la relación positiva entre el tamaño de la muestra (*VI*) y el poder estadístico (*VD*).

Figura 10.8 **Planeación de la Muestra con el *d* de Cohen y el Poder Estadístico con una Cola.**

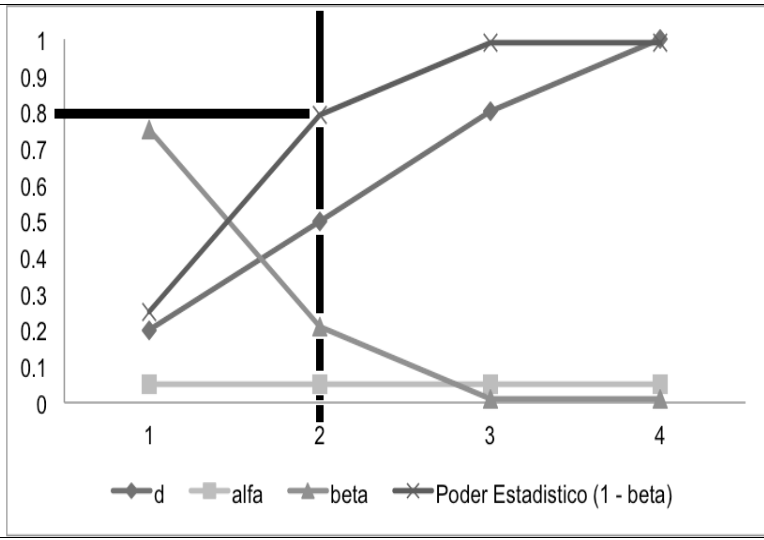
e. Comparación de Una y Dos Colas con el Coeficiente Beta y el Poder Estadístico.

La Figura 10.11 muestra una comparación entre un test que involucra una cola y dos colas. Para ilustrar este ejemplo de comparación se usaron los promedios con un test *t* de estudiante de grupos independientes (*n* = 50) con un alfa fijo de .05 para ambos test. En el *plano cartesiano* de esta figura, aparece en el eje de la *x* los cuatro escenarios (i.e., 1, 2, 3, y 4, donde *d* = 0.20, *d* = 0.50, *d* = 0.80, y *d* = 1.00, respectivamente), y el eje de la *y* se representan las probabilidades (0 probabilidad a 1 que es el 100% de probabilidad). El propósito de la Figura 10.11 es mostrar como al incrementar el *d*, el coeficiente beta disminuye y el poder estadístico aumenta, pero este último aumenta más rápidamente para el test de una cola debido a que el alfa se concentra en una sola cola y, por ello, disminuye el área del coeficiente beta. Al disminuir el coeficiente beta, el poder estadístico aumenta (Figura 10.11).

En la Figura 10.9, la Grafica del Test de una Cola muestra que la línea que representa al *d* aumenta de 0.20 a 1.00 (i.e., del Escenario 1 al 4), manteniendo al $\alpha = .05$ (línea paralela al eje de la *x*). Este aumento causa que el coeficiente beta disminuya, y el poder estadístico incremente. Este poder estadístico alcanza el mínimo de .80 recomendado por Cohen (1988) en el Escenario 2 cuando el $d = 0.50$. Con el fin de señalar el poder estadístico del 80%, estos puntos están señalados por una línea perpendicular al eje de las *x* (i.e., a la altura del número 2: Escenario 2), y otra también perpendicular al eje de la *y* (i.e., probabilidad de .80).

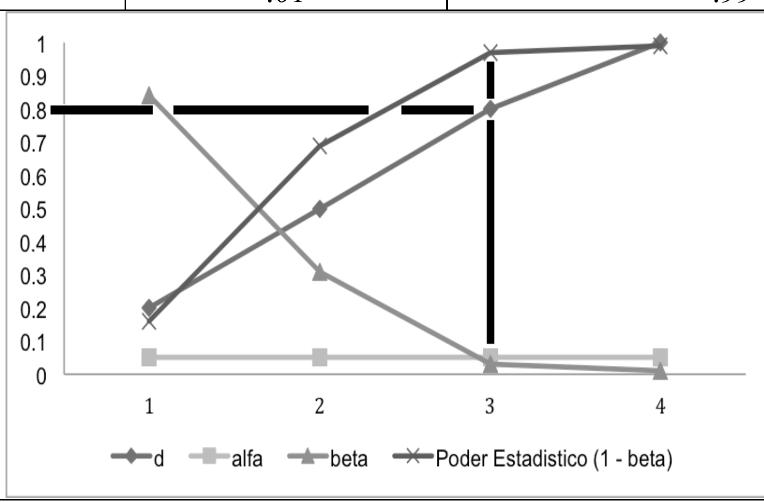
Por otro lado, la Grafica del Test de dos Colas muestra que no se alcanza el poder estadístico de .80 cuando el $d = 0.50$ en el Escenario 2 como en el test de una cola, cereteris paribus (Figura 10.11). Es decir, el Escenario 2 muestra que el poder estadístico es de aproximadamente .69 cuando $d = 0.50$. Cuando el $d = .80$, el poder estadístico sobrepasa al mínimo antes mencionado porque es .97. Lo anterior indica que el poder estadístico de .80 debe de estar entre el de $d = 0.50$ (Escenario 2) y $d = .80$ (Escenario 3). El Escenario 3 también se señala con líneas perpendiculares al eje de la *x* y de la *y* para evidenciar lo explicado en este párrafo.

En resumen y como recomendación, una investigadora o investigador conseguirá más poder estadístico cuando puede precisar la dirección de su hipótesis alternativa con una cola, céteris páribus: i.e., ya sea $\bar{x}_1 < \bar{x}_2$ o $\bar{x}_1 > \bar{x}_2$. En la investigación educativa seria relativamente sencillo el precisar que un pre and post test tendrán una diferencia después de algún tratamiento educativo. Es decir, se esperaría que el post-test tenga puntajes más altos porque el tratamiento debió de haber mejorado algo. Por lo tanto, habría que especificar en la alternativa lo anterior, y, con ello, tener más poder estadístico, céteris páribus: H_A: $\bar{x}_{\textit{pre-test}} < \bar{x}_{\textit{post-test}}$. Además, como el alfa se concentra en una sola cola se tendrá más probabilidades de rechazar la hipótesis nula que cuando se tienen dos colas. Por ejemplo, se elige un alfa = .05 y se pone a prueba con un test de una cola y de dos colas. La probabilidad calculada es igual .04, la cual sería estadísticamente significativa para una cola (probabilidad $_{calculada}$ = .04 < alfa .05). Sin embargo, no lo seria para el test de dos colas (probabilidad $_{calculada}$ = .04 > alfa / 2 = .025).



Grafica del Test de una Cola; $n = 50$.

Escenario y *d* de Cohen	Beta	Poder Estadistico
Escenario 1: 0.20	.75	.25
Escenario 2: 0.50	**.20**	**.80**
Escenario 3: 0.80	.01	.99
Escenario 4: 1.00	.01	.99



Grafica del Test de dos Colas

Escenario y *d* de Cohen	Beta	Poder Estadistico
Escenario 1: 0.20	.84	.16
Escenario 2: 0.50	.31	.69
Escenario 3: 0.80	**.03**	**.97**
Escenario 4: 1.00	.01	.99

Nota: El eje de la *x* (linea orizontal del plano cartesiano) representa los 4 Escenarios: 1, 2, 3, y 4. La y (linea vertical) representa la probabilidad que va de 0 a 1 (100%).

Figura 10.9 **Una y dos Colas, *d* de Cohen, Alfa, Beta, y Poder Estadístico.**

Preguntas para Reflexionar

1.- ¿Cuáles son las consecuencias de no tener un poder estadístico del 80%?

2.- ¿Cómo se podría llegar a tener un poder estadístico del 80%?

3.- ¿Cuáles son las implicaciones de aumentar el alfa?

4.- ¿Por qué es mejor usar un test de una sola cola?

5.- ¿Cuándo es más apropiado usar un test de dos colas?

Opinión del Autor

Para algunos estudiantes de investigación educativa resulta complicado el entender como el poder estadístico tiene una relación con todas estas variables simultáneamente: el alfa, el tamaño de la muestra y el *d* de Cohen. Cuando uno de estas variables cambia, también lo hace el poder estadístico. Lo más importante de estas relaciones es con el tamaño de la muestra porque es algo que tanto el estudiante como el investigador o investigadora puede manipular hasta cierto punto. Aunque el alfa también se puede manipular, el hacerlo puede causar sospechas a los lectores porque el investigador o la investigadora pueden estar buscando incrementar su poder estadístico sin importar el caer en el Error tipo I. La variable que no se puede manipular es el *d* de Cohen que surge de los análisis de los datos. Sin embargo, el *d* de Cohen se puede modificar para estimar el tamaño de la muestra que se usara en un estudio con un 80% de poder estadístico. Para ello, se pueden asumir tres escenarios: malo con un $d = 0.20$; uno regular con un $d = 0.50$; y uno bueno con un $d = 0.80$.

Mi recomendación es que el lector juegue con el software del *Gpower* como yo lo hice en estas secciones. Esto con el propósito de familiarizarse con el cambio en el poder estadístico a partir de la variación en las otras variables.

Capítulo 11. Tamaño del Efecto: El *d* de Cohen.

Existen muchos tamaños del efecto que también son conocidos como el efecto práctico (e.g., eta cuadrado, *V* de Cramer, *f* de Cohen, la Correlación Producto Momento de Pearson [*r*], el Coeficiente de Determinación [r^2], el *d* de Cohen, etc.). Posiblemente uno de los más simples de calcular y de interpretar es el *d* de Cohen. En muchas ocasiones, aparece en la literatura que las y los investigadores confunden el efecto estadístico con el efecto práctico. En concreto, creen que el tamaño de la probabilidad calculada es la magnitud de la diferencia entre la comparación de dos promedios. Piensan que necesariamente entre más pequeña sea la probabilidad calculada (i.e., $p \leq .01$), mayor será el efecto práctico. Esto es parcialmente cierto porque un *d* de Cohen grande se suele asociar con una probabilidad calculada pequeña cuando se tiene una muestra grande (e.g., $n > 100$). Sin embargo, el efecto estadístico es sensible al tamaño de la muestra y el *d* de Cohen no lo es en su fórmula. Esto quiere decir que, si se logra tener muestras grandes, se puede encontrar un efecto estadístico (probabilidad calculada < alfa), pero no necesariamente se obtendrá un *d* de Cohen grande asociado con este efecto estadístico (cf. Cohen, 1988). Puede suceder lo contrario con muestras pequeñas: un *d* de Cohen grande sin efecto estadístico (probabilidad calculada > alfa).

Cohen (1988, p. 8) explico que con un test de significancia estadística (i.e., *Test de Significancia Estadística de la Hipótesis Nula*) se puede observar si un fenómeno existe cuando la hipótesis nula es falsa o el fenómeno está ausente cuando la hipótesis nula es cierta. El primer escenario que explicó Cohen (1988) es cuando un investigador o investigadora encuentra lo siguiente:

H_0: $\bar{x}_1 = \bar{x}_2$ con (probabilidad calculada > alfa).

Entonces, se falla en rechazar la hipótesis nula.

Por otro lado, esto es lo que sucede a nivel de poblaciones: $\mu_1 \neq \mu_2$.

Por lo tanto, la H_0 era falsa, y, como resultado, el fenómeno existe.

Para ilustrar la contraparte de la nula existencia de un fenómeno, es que una o un investigador encuentre lo siguiente:

H_0: $\bar{x}_1 = \bar{x}_2$ (probabilidad calculada > alfa).

Entonces, se falla en rechazar la hipótesis nula.

Cuando esto sucede a nivel poblaciones: $\mu_1 = \mu_2$.

Por lo tanto, la H_0 era cierta, y, entonces, el fenómeno no existe.

Sin embargo, difícilmente un investigador o investigadora podrá saber si una hipótesis nula es falsa porque no se suele tener acceso a toda la población. Se puede tener cierta evidencia de que es falsa cuando se rechaza la nula, pero no se puede estar seguro de que sea falsa. Además de este test de efecto estadístico, Cohen (1988) describió una serie de efectos prácticos para complementar estos primeros. Un efecto fue definido por Cumming (2013, p. 440) como, "Algo en lo cual podríamos estar interesados."

Contextualizando esta definición y volviendo al tema de la comparación entre un promedio de alguna materia escolar de dos poblaciones (niñas vs. niños), el efecto sería el cambio o diferencia que produce el sexo en estos promedios. Se puede decir que el efecto es la relación de causa y efecto entre una variable independiente y una dependiente: Variable independiente → Variable Dependiente (la flecha indica causalidad). Para más información al respecto, ver al diccionario de Psicología del APA (*APA Dictionary of Psychology*, 2015, p. 352).

De hecho, Cohen (1988) creó un coeficiente para medir la magnitud del cambio entre dos promedios: *d* de Cohen. Este coeficiente es conocido como uno de los efectos prácticos (para más información sobre efectos prácticos, ver a Cohen, 1992). Al igual que el efecto estadístico, Cohen (1988) explico que, si el *d* de Cohen existiera en las poblaciones, entonces el promedio de una no sería igual al promedio de la otra:

i.e., ($\mu_1 \neq \mu_2$) y la hipótesis nula seria falsa para las muestras ($\bar{x}_1 = \bar{x}_2$).

El tamaño del efecto (*effect size*) fue también definido por Cumming (2013, p. 440) como, "La cantidad de algo que podría ser de interés. Es el tamaño de un efecto." Usando ejemplo de las niñas vs. los niños, el tamaño de efecto sería la magnitud de la diferencia entre ambos promedios. Esta diferencia podría ir del 0 al infinito. Para estimar esta diferencia, se puede usar el *d* de Cohen. Cumming (2013, p. 439) definió al *d* como, "Un tamaño de efecto estandarizado expresado en unidades de acuerdo con la desviación estándar. En muchas ocasiones, puede ser considerado como un tipo de valor *z*." Cohen (1988, pp. 25-26) dio unos tamaños de efecto como referencia general (i.e., cuando no existe en la literatura alguna referencia para determinar el tamaño de un efecto): 0.20 = pequeño; 0.50 = mediano; y 0.80 = grande. La Ecuación (11.1) para estimar el *d* para las poblaciones es:

$$\delta = \frac{(\mu_1 - \mu_2)}{\sigma} \qquad \text{Ecuación (11.1)}$$

δ = el coeficiente del *d* de Cohen para las poblaciones, y se pronuncia delta.

μ_1 = promedio de la población 1.

μ_2 = promedio de la población 2.

σ = desviación estándar de una población porque se asume que ambas desviaciones de las poblaciones son iguales.

Cuando se calcula el *d* de Cohen para dos muestras (Ecuación 11.2):

$$d = \frac{(\bar{x}_1 - \bar{x}_2)}{SD^*} \qquad \text{Ecuación (11.2)}$$

$\bar{x}_1$ = promedio de la muestra 1.

$\bar{x}_2$ = promedio de la muestra 2.

SD^* = la desviación estándar de una muestra cuando es igual a la otra, pero si son diferentes $SD_1 + SD_2 / 2$ cuando viene de muestras equivalentes en número de participantes: e.g., n =25 para la muestra 1 y n = 25 para la muestra 2.

SD_1= Desviación estándar de la muestra 1.

SD_2 = Desviación estándar de la muestra 1.

Nota: Este coeficiente no se puede interpretar cuando tiene un valor negativo, así que se toma con su valor positivo para ser interpretado. De hecho, el valor del *d* es absoluto: i.e., |*d*|. Existen otras fórmulas para calcular el *d*, y, para estimar dichas formulas, se recomienda a Thalheimer y Cook (2002).

Para ilustrar cuatro tamaños del tamaño del efecto (i.e. *d* = 0; *d* = 0.20; *d* = 0.50; *d* = 0.80), se muestra la Figura 11.1. Esta serie de ejemplos, se llevaron a cabo en el Software de *Gpower* usando un test *t* de estudiante de muestras independientes con un alfa = .05, *n* = 50, y grupos equivalentes con un test *post hoc* para calcular el poder estadístico. Dos curvas normales representan la distribución de los datos de las dos muestras, las cuales tienen la misma varianza. Para la siguiente serie de ejemplos, se muestra el uso de la fórmula del *d* (i.e., $d = [\bar{x}_1 - \bar{x}_2] / SD$). Asimismo, la Figura 11.1 ilustra con curvas normales esta serie de ejemplos del uso de la fórmula del *d*:

Cuando el *d* = 0, las dos distribuciones de las dos muestras se empalman completamente. En el siguiente ejemplo, se les dan valores a los promedios para ilustrar:

d: 0 = (100 - 100) / 100

Cuando el *d* = 0.20, en este caso los promedios se alejan con un 1 / 5 de *SD*. Este efecto pequeño, se puede deber a que en nuevas áreas de estudio no se tiene un buen control o la mediación es inexacta o ambos (Cohen, 1988, p. 25):

d: 0.20 = (100 - 80) / 100

Cuando el *d* = 0.50, en este caso los promedios se alejan con un 1 / 2 de *SD*. Este efecto es mediano, y Cohen (1988, p. 26) explico que este tamaño de efecto es visible a simple vista. Como un ejemplo, él detallo que se puede notar la diferencia del nivel de inteligencia entre trabajadores manuales y profesionistas:

d: 0.50 = (100 - 50) / 100

Cuando el *d* = 0.80, en este caso los promedios se alejan con un 4 / 5 de *SD*. Este efecto grande, Cohen (1988, p. 26) mostró un ejemplo en el cual una diferencia de efecto grande sucedería entre la inteligencia de una persona con un doctorado y un estudiante de primer año de universidad.

d: 0.50 = (100 - 20) / 100

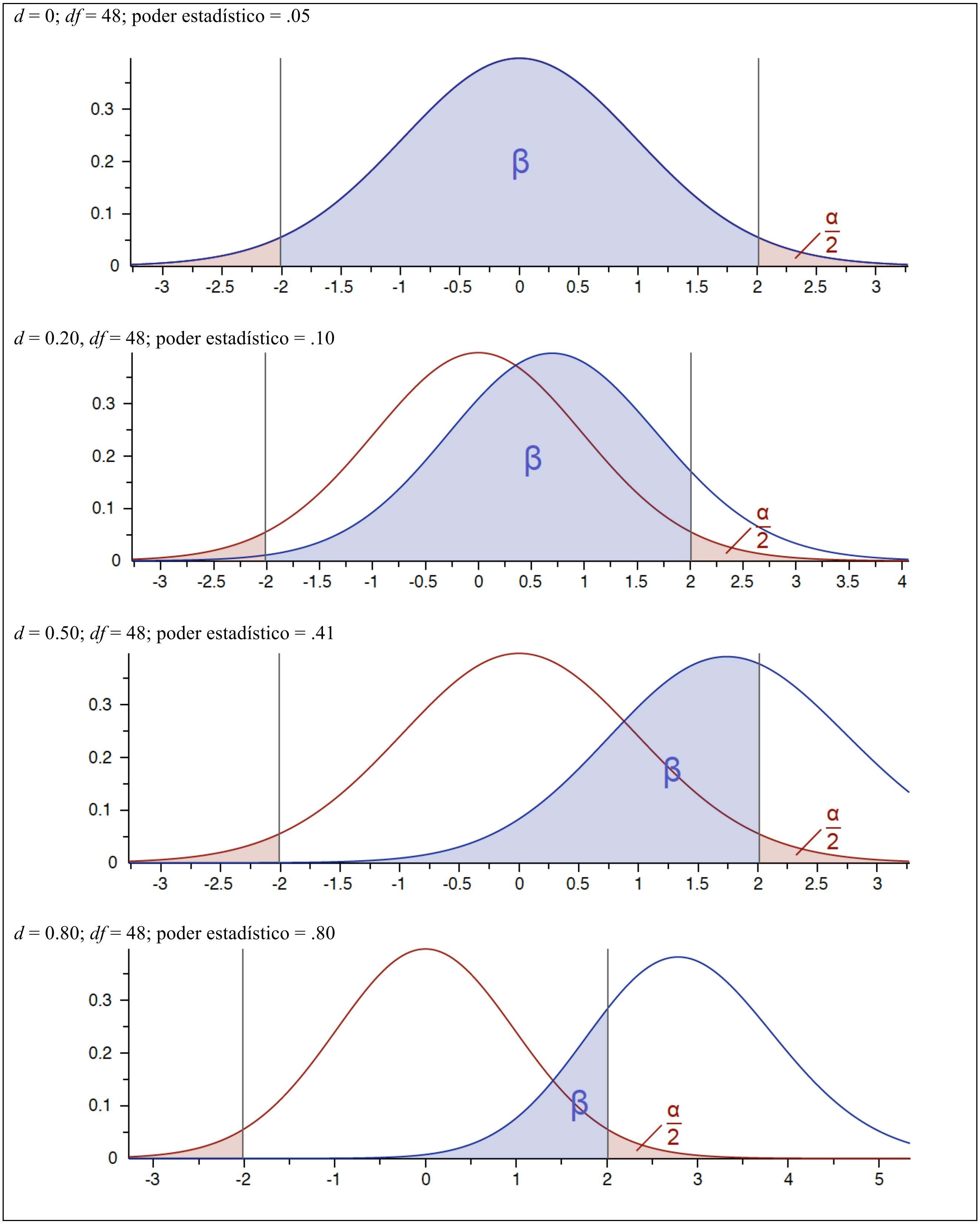


Figura 11.1 **Coeficiente del *d* de Cohen.**

Puede pasar que al comparar dos grupos independientes con el *d* estos tengan diferentes tamaños de muestra. Por lo tanto, una formula diferente a la *d* de Cohen (i.e., d_s) es necesaria y esta es para la diferencia *estandarizada* entre grupos de muestras de observaciones independientes (Ecuación 11.3):

$$d_s = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 - n_2 - 2}}} \qquad \text{Ecuación (11.3)}$$

$\bar{x}_1$ = Promedio de la muestra 1.

$\bar{x}_2$ = Promedio de la muestra 2.

n_1 = Tamaño de la muestra 1.

SD_1^2 = Desviación estándar al cuadrado de la muestra 1.

n_2 = Tamaño de la muestra 2.

SD_2^2 = Desviación estándar al cuadrado de la muestra 2.

Existen varios artículos y libros que han hablado extensivamente de los efectos prácticos como el *d* de Cohen, entre otros. Para un repaso rápido del *d* de Cohen y otros efectos, se recomienda los artículos de Cohen (1992) y Lakens (2013). Para un tratamiento más amplio de los efectos prácticos, se recomienda a libro de Cohen (1988), especialmente si se estudia la relación del *d* con el poder estadístico. Asimismo, si se quiere calcular un intervalo de confianza para el *d*, se recomienda el libro de Cumming (2013). Un resumen de los tamaños del *d* de Cohen, según Cohen (1988): 0.20 pequeño; .0.50 mediano; y 0.80 grande; Para aumentar las categorías del *d*, Sawilowsky (2009) agregó: 0.01 muy pequeño; 1.20 muy grande; y 2.0 enorme.

Preguntas para Reflexionar

1.- ¿Es el *d* de Cohen usable en todas las áreas del conocimiento?

2.- ¿Qué otros tamaños del *d* de Cohen deberían de ser clasificados además de los anteriores?

3.- ¿Cómo definir un *d* calculado que se encuentre exactamente entre dos categorías diferentes como mediano y grande?

4.- ¿Debería de haber intervalos en las categorías para poder precisar el tamaño de acuerdo con algún adjetivo?

Opinión de Autor

La elegancia del *d* de Cohen no deja de sorprenderme por su simplicidad dado que el efecto estadístico con sus tipos de errores, alfa, beta y poder estadístico obscurecen su interpretación en algunos casos. Muchas veces son demasiadas variables para el efecto estadístico. En contraparte, el *d* es fácil de entender porque es una diferencia entre promedios dividida por la desviación estándar, céteris páribus. Mis estudiantes suelen entender este coeficiente, y esto es especialmente cuando hay un diseño experimental donde hay que comparar grupos antes del tratamiento y después de este con el *d*. Es un caso diferente cuando se usan coeficientes que tratan con varianzas explicas y varianzas sin explicar. Mi recomendación es usar el *d* como uno de los elementos más importantes en la comparación de grupos. De hecho, Cumming (2013) y la Asociación Americana de Psicología, entre muchos otros más, han explicado la importancia de reportar un tamaño del efecto como el *d*.

Conclusión

De nuevo y desde un punto de vista teórico, el objetivo del presente libro es mostrar cómo llevar a cabo test de significancia estadística con sus implicaciones, explicar principios básicos de muestreo, estimar intervalos de confianza para el promedio, y calcular el tamaño de un efecto práctico: *d* de Cohen. Específicamente, los siguientes párrafos explican este objetivo.

Del contenido de este libro, se puede concluir que la curva normal es un modelo que sirve para representar fenómenos de la naturaleza, pero también fenómenos de la educación como el aprendizaje, entre otros. Este modelo se puede usar cuando los datos se distribuyen de una manera aproximadamente normal. El Teorema de Tendencia Central muestra que una serie de muestras constituirán una distribución aproximadamente normal. La curva normal estandariza es la versión operacional de la anterior. En esta, se pueden calcular probabilidades de acuerdo con la distancia entre un valor z y el promedio del z que es 0.

Con las estadísticas inferenciales, se trata de tomar una o varias muestras para calcular una estadística de interés: e.g., promedio. La estadística representa al parámetro de la población. Sin embargo, hay que considerar un error que es la diferencia entre una estadística y un parámetro, y, para ello, se usan los intervalos de confianza que estiman este error y capturan supuestamente el parámetro verdadero de la población.

Las estadísticas paramétricas se usan cuando se asume que una muestra viene de una población de interés. Los datos de la muestra deben de estar aproximadamente distribuidos normalmente para poder usar las estadísticas paramétricas. Por otro lado, las estadísticas no-paramétricas no vienen de poblaciones, sino que se analizan los datos en sí mismos. Entonces, los análisis no paramétricos no necesitan tener una distribución normal.

Las variables se pueden dividir en independientes y dependientes cuando se espera un efecto de una sobre la otra o por lo menos una asociación entre ambas. Dependiendo del diseño de la investigación (i.e., experimental y no-experimental), el investigador o investigadora tiene control o algún tipo de control sobre la variable independiente.

Las hipótesis se derivan de las preguntas de investigación. Una de las hipótesis (nula) afirma que la diferencia entre dos grupos en cuanto a la variable dependientes es cero. En contraparte, la hipótesis alternativa especifica lo contario: la diferencia no es cero.

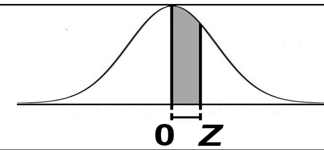
Los errores se cometen cuando uno compara los promedios de muestras e infiere que hay una diferencia, por un lado, o no hay diferencias, y se generaliza a las poblaciones en las cuales puede que exista esa diferencia o no. En detalle, se puede encontrar una diferencia entre las muestras cuando esta no existe entre las poblaciones (Error Tipo I). El otro error viene cuando no se encuentra diferencia entre las muestras, pero si hay diferencia entre las poblaciones (Error Tipo II).

Se asume que un fenómeno existe si el promedio de una población y otra son diferentes. Este es el efecto estadístico. Un ejemplo es que se definan a dos naciones como dos poblaciones diferentes entre sí. Al comparar sus niveles de aprendizaje en alguna materia escolar se encuentra que sus promedios son diferentes. De lo anterior, se puede inferir que existe un efecto de las naciones en los niveles de aprendizaje. También, se asume que un fenómeno existe si un parámetro como el promedio de dos poblaciones es diferente en cierta dirección previamente identificada. Un ejemplo heurístico es que dos naciones tienen los mismos niveles de aprendizaje en alguna materia escolar. Se ha identificado empíricamente que las tutorías mejoran el aprendizaje. Uno de los países implementa las tutorías. Luego de un tiempo, se esperaría que el país que implemento las tutorías tuviera más altos niveles de aprendizaje que el otro. Esto último seria la dirección en la que se espera la diferencia.

En la probabilidad y estadísticas existen varias variables que interactúan entre sí. Unas de estas variables son manipulables como el tamaño de la muestra y la elección de elegir una o dos colas para el test de significancia. Entre más grande sea la muestra se tendrá más poder estadístico para rechazar la hipótesis nula cuando sea falsa, céteris páribus. De una manera similar, si se elige una cola se tendrá más poder estadístico que cuando se eligen dos colas, y se tendrá más posibilidad de encontrar resultados significativos ya que el alfa (nivel de significancia) estará concentrado en una sola cola.

Un tamaño de efecto es el *d* de Cohen que sirve para medir la magnitud de la diferencia de dos promedios al producir un coeficiente estandarizado que se ha clasificado como: pequeño, mediano y grande. Con estos puntos de referencia, cuando no hay otros en la literatura, se puede evaluar con el *d* en alguna investigación educativa.

Apéndice

Valores z					
Valor | z_i |	Área entre el promedio y z_i	Valor | z_i |	Área entre el promedio y z_i	Valor | z_i |	Área entre el promedio y z_i
.00	.0000	.50	.1915	1.00	.3413
.01	.0040	.51	.1950	1.01	.3438
.02	.0080	.52	.1985	1.02	.3461
.03	.0120	.53	.2019	1.03	.3485
.04	.0160	.54	.2054	1.04	.3508
.05	.0199	.55	.2088	1.05	.3531
.06	.0239	.56	.2123	1.06	.3554
.07	.0279	.57	.2157	1.07	.3577
.08	.0319	.58	.2190	1.08	.3599
.09	.0359	.59	.2224	1.09	.3621
.10	.0398	.60	.2257	1.10	.3643
.11	.0438	.61	.2291	1.11	3665
.12	.0478	.62	.2324	1.12	.3686
.13	.0517	.63	.2357	1.13	.3708
.14	.0557	.64	.2389	1.14	.3729
.15	.0596	.65	.2422	1.15	.3749
.16	.0636	.66	.2454	1.16	.3770
.17	.0675	.67	.2486	1.17	.3790
.18	.0714	.68	.2517	1.18	.3810
.19	.0753	.69	.2549	1.19	.3830
.20	.0793	.70	.2580	1.20	.3849
21	.0832	.71	.2611	1.21	.3869
.22	.0871	.72	.2642	1.22	.3888
.23	.0910	.73	.2673	1.23	.3907
.24	.0948	.74	.2703	1.24	.3925
.25	.0987	.75	.2734	1.25	.3944
.26	.1026	.76	.2764	1.26	.3962
.27	.1064	.77	.2794	1.27	.3980
.28	.1103	.78	.2823	1.28	.3997
.29	.1141	.79	.2852	1.29	.4015
.30	.1179	.80	.2881	1.30	.4032
.31	.1217	.81	.2910	1.31	.4049
.32	.1255	.82	.2939	1.32	.4066
.33	.1293	.83	.2967	1.33	.4082
.34	.1331	.84	.2995	1.34	.4099
.35	.1368	.85	.3023	1.35	.4115
.36	.1406	.86	.3051	1.36	.4131
.37	.1443	.87	.3178	1.37	.4147
.38	.1480	.88	.3106	1.38	.4162
.39	.1517	.89	.3133	1.39	.4177
.40	.1554	.90	.3159	1.40	.4192
.41	.1591	.91	.3186	1.41	.4207
.42	.1628	.92	.3212	1.42	.4222
.43	.1664	.93	.3238	1.43	.4236
.44	.1700	.94	.3264	1.44	.4251
.45	.1736	.95	.3289	1.45	.4265
.46	.1772	.96	.3315	1.46	.4279
.47	.1808	.97	.3340	1.47	.4292
.48	.1844	.98	.3365	1.48	.4306
.49	.1879	.99	.3389	1.49	.4319
.50	.1915	1.00	.3413	1.50	.4332

Continuación

Valor | z_i |	Área entre el promedio y z_i	Valor | z_i |	Área entre el promedio y z_i	Valor | z_i |	Área entre el promedio y z_i
1.50	.4332	2.00	.4772	2.50	.4938
1.51	.4345	2.01	.4778	2.51	.4940
1.52	.4357	2.02	.4783	2.52	.4941
1.53	.4370	2.03	.4788	2.53	.4943
1.54	.4382	2.04	.4793	2.54	.4945
1.55	.4394	2.05	.4798	2.55	.4946
1.56	.4406	2.06	.4803	2.56	.4948
1.57	.4418	2.07	.4808	2.57	.4949
1.58	.4429	2.08	.4812	2.58	.4951
1.59	.4441	2.09	.4817	2.59	.4952
1.60	.4452	2.10	.4821	2.60	.4953
1.61	.4463	2.11	.4826	2.61	.4955
1.62	.4474	2.12	.4830	2.62	.4956
1.63	.4484	2.13	.4834	2.63	.4957
1.64	.4495	2.14	.4838	2.64	.4959
1.65	.4505	2.15	.4842	2.65	.4960
1.66	.4515	2.16	.4846	2.66	.4961
1.67	.4525	2.17	.4850	2.67	.4962
1.68	.4535	2.18	.4854	2.68	.4963
1.69	.4545	2.19	.4857	2.69	.4964
1.70	.4554	2.20	.4861	2.70	.4965
1.71	.4564	2.21	.4864	2.71	.4966
1.72	.4573	2.22	.4868	2.72	.4967
1.73	.4582	2.23	.4871	2.73	.4968
1.74	.4591	2.24	.4875	2.74	.4969
1.75	.4599	2.25	.4878	2.75	.4970
1.76	.4608	2.26	.4881	2.76	.4971
1.77	.4616	2.27	.4884	2.77	.4972
1.78	.4625	2.28	.4887	2.78	.4973
1.79	.4633	2.29	.4890	2.79	.4974
1.80	.4641	2.30	.4893	2.80	.4974
1.81	.4649	2.31	.4896	2.81	.4975
1.82	.4656	2.32	.4898	2.82	.49.76
1.83	.4664	2.33	.4901	2.83	.4977
1.84	.4671	2.34	.4904	2.84	.4977
1.85	.4678	2.35	.4906	2.85	.4978
1.86	.4686	2.36	.4909	2.86	.4979
1.87	.4693	2.37	.4911	2.87	.4979
1.88	.4699	2.38	.4913	2.88	.4980
1.89	.4706	2.39	.4916	2.89	.4981
1.90	.4713	2.40	.4918	2.90	.4981
1.91	.4719	2.41	.4920	2.91	.4982
1.92	.4726	2.42	.4922	2.92	.4982
1.93	.4732	2.43	.4925	2.93	.4983
1.94	.4738	2.44	.4927	2.94	.4984
1.95	.4744	2.45	.4929	2.95	.4984
1.96	.4750	2.46	.4931	2.96	.4985
1.97	.4756	2.47	.4932	2.97	.4985
1.98	.4761	2.48	.4934	2.98	.4986
1.99	.4767	2.49	.4936	2.99	.4986
2.00	.4772	2.50	.4938	3.00	.4987

Referencias

American Psychological Association. (2015). *APA Dictionary of Psychology* (2ª ed.). Washington, DC: American Psychological Association.

Berenson, M.L., Levine, D.M., y Krehbiel, T.C. (2012). *Basic Business Statistics: Concepts and Applications* (12ª ed.). Nueva York: Pearson.

Berk, L.E. (2007). *Development Through the Lifespan* (4ª ed.) Nueva York: Pearson.

Box, G.E.P. y Tiao, G.C. (1973). *Bayesian Inferences in Statistical Analysis.* Massachusetts: Addison-Wesley Publishing Company.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2ª ed.). Nueva York: Psychology Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Crocker, L. y Algina, J. (2006). *Introduction to Classical & Modern Test Theory*. Mason, Ohio: Cengage Learning.

Cumming, G. (2013). *Understanding the New Statistics: Effect sizes, confidence intervals, and meta-analysis.* Nueva York: Routledge.

DeVellis, R.F. (2016). *Scale Development: Theory and Applications* (4ª ed.). Los Ángeles: Sage.

ElAtia, S., Ipperciel, D., y Zaiane, O.R. (2016). *Data Mining and Learning Analytics: Applications in Educational Research.* Nueva Jersey: Wiley.

Embretson, S.E. y Reise, S.P. (2000). *Item Response Theory for Psychologists*. Nueva Jersey: Lawrence Erlbaum Associates Publishers.

Enders, G.K. (2010). *Applied Missing Data Analysis.* Nueva York: The Guilford Press.

Ercikan, K. y Roth, W. M. (2006). What good is polarizing research into qualitative and quantitative? *Educational researcher*, *35*(5), 14-23.

Fischer, H. (2010). *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Nueva York: Springer.

Hernández-Sampieri, R., Fernández–Collado, C., y Baptista-Lucio, M. P. (2014). *Metodología de la Investigación* (6ª ed.). México, D.F.: McGraw Hill.

Hinkle, D.E., Wiersma, W., y Jurs, S.G. (2003). *Applied Statistics for Behavioral Sciences* (5ª ed.). Nueva York: Houghton Mifflin.

Hurley, P.J. (2014). *A concise Introduction to Logic* (12ª ed.). San Diego: Cengage Learning.

Johnson, R. (1990). *Estadística Elemental*. México: Editorial Trillas.

Kotz, S. (Ed.) (2006). *Encyclopedia of Statistical Sciences: 16 volumes* (2ª ed.). Nueva Jersey: Wiley-Interscience.

Kuder, G.F. y Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.

Kutner, M.H., Nachtsheim, C.J., y Neter, J. (2004). *Applied Linear Regression Models* (4a ed.). Boston, Massachusetts: McGraw-Hill.

Lane, D. M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2014). *Introduction to statistics*. Rice Univ., Houston*, TX*: David Lane.

Maxwell, S.E. y Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (2ª ed.). Mahwah, Nueva Jersey: Lawrence Erlbaum Associates, Publishers.

McMillan, J.H. y Schumacher, S. (2007). Investigación Educativa (5ª ed.). México: Pearson.

Ponce, H.F. (aceptado y en prensa). Evaluación del proceso de validación de tests psicométricos a través del Análisis Exploratorio de Factores. *Revista Internacional de Medición y Evaluación.*

Ponce, H.F. (aceptado y en prensa). ¿Qué tan apropiadamente reportaron los autores el Coeficiente del Alfa de Cronbach? *Revista Internacional de Medición y Evaluación.*

Ponce, H.F., Domínguez, C.T., Arriaga, M. (2016). La importancia de la investigación en la educación especial. *Noesis, 25*(50), 216-242.

Ponce, H.F., Soto, M., y Solorio, M.M. (2017, abril). ¿A qué carreras universitarias van los mejores puntajes de un examen de admisión? *Congreso de la AcademiaJournals.com, 9*(2), 716-721.

Ross, S.M. (1997). *Introduction to Probability Models* (6ª ed.). Nueva York: Academic Press.

Salkind, N.J. (Ed.). (2007). *Encyclopedia of Measurement and Statistics* (Volumes 1, 2, 3). Thousand Oaks, California: Sage.

Salkind, N.J. (2011). *Statistics for People Who Think They Hate Statistics* (4ª ed.). Los Angeles: Sage.

Sawilowsky, S.S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), 597-599.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., y Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs. *Washington, DC: American Educational Research Association*.

Schumacker, R. y Tomek, S. (2013). *Understanding Statistics: Using R.* Nueva York: Springer.

Shavelson, R.J. y Webb, N.M. (1991). *Generalizability theory: A primer*. Thousand Oaks, California. Sage.

Siegel, J. y Castellan, N.J. (1995). *Estadística no paramétrica a las ciencias de la conducta*. México: Trillas.

Sudman, S. (1976). *Applied Sampling.* Nueva York: Academic Press.

Tabachnick, B.G. y Fidell, L.S. (2012). *Using Multivariate Statistics* (6ª ed.). Nueva York: Pearson.

Thalheimer, W. y Cook, S. (2002). *How to calculate effect sizes from published research articles: A simplified methodology.* Recuperado el 4 de septiembre, 2018, de www.bwgriffin.com/gsu/courses/edur9131/content/Effect_Sizes_pdf5.pdf

Thompson, B. (Ed.) (2003). *Score Reliability: Contemporary Thinking on Reliability Issues*. Thousand Oaks, California: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. En J.W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 246-262). Thousand Oaks, California: Sage.

Wechsler, D. (1997). *The Wechsler Adult Intelligence Scale* (3ª ed.). San Antonio, TX: The Psychological Corporation.

Yu, C.H. (2006). *Philosophical Foundations of Quantitative Research Methodology*. Lanham, Oxford: University Press of America.