

CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometrics for Evaluation

3 Experimentos II

Disclaimer: I AM NOT the original intellectual author of the material presented in these notes. The content is STRONGLY based on a combination of lecture notes (from Aurora Ramirez), textbook references, and personal annotations for learning purposes. Any errors or omissions are entirely my own responsibility.

Índice

Sesgo y precisión	3
Evaluamos el efecto de traer tutores a las escuelas	3
Diferencia entre las medias muestrales	4
¿Qué pasaría si corriéramos un segundo experimento?	5
Múltiples experimentos: una distribución de las estimaciones	6
Distribución de las estimaciones si el verdadero efecto = β	7
Distribución de las estimaciones si el verdadero efecto = 0	8
Dos distribuciones bajo dos hipótesis distintas	9
Si corremos un solo experimento	9
¿Nuestra estimación provino de H_β o H_0 ?	10
Nivel de significancia del 5 %	11
Valor crítico	11
¿Es β significativamente diferente de 0, al 5 % de significancia?	12
Prueba de hipótesis	13
Prueba de hipótesis (evaluación de programas)	13
¿Qué es el nivel de significancia?	14
¿Qué es poder estadístico?	14
Cuatro resultados de las pruebas de hipótesis	15
¿Cuántas veces se rechazaría la hipótesis nula, si es que H_β es cierta?	15
¿Cuáles son los determinantes del poder estadístico?	16
Poder estadístico: principales determinantes	17
Tamaño del efecto = $1 \times$ error estándar	17
Tamaño del efecto = $1 \times$ error estándar	18
Poder igual a 26 % si el verdadero impacto es $1 \times$ error estándar	19

Tamaño del efecto = $3 \times$ error estándar. Poder = 91 %	20
Tamaño del efecto y adopción (take-up)	21
Al incrementar el tamaño de la muestra...	22
Tamaño de la muestra	23
¿Cómo un incremento en la varianza de la población afecta las curvas de distribución de las estimaciones?	24
Varianza	24
Varianza y poder: intuición	25
Proporción de la muestra en T vs. C	27
¿Qué pasa si la división de la muestra no es del 50:50?	27
¿Qué tanto es muy desequilibrado?	28
Ecuación de poder. Efecto Mínimo Detectable	29
Diseño agrupado. Definición	30
Baja correlación dentro de la agrupación o clúster (ICC, ρ)	30
Alta correlación dentro de la agrupación o clúster (ρ)	31
¿Cómo ρ (ICC) influye en el poder estadístico?	31
Calcular poder con Stata	32

Sesgo y precisión

- La **precisión** aumenta con el **tamaño de muestra**: más observaciones generan estimaciones menos dispersas.
- El **sesgo** se reduce con la **aleatorización**: asegura que las estimaciones estén centradas en la verdad.
- La calidad de un experimento se mide por la combinación de **bajo sesgo** y **alta precisión**.

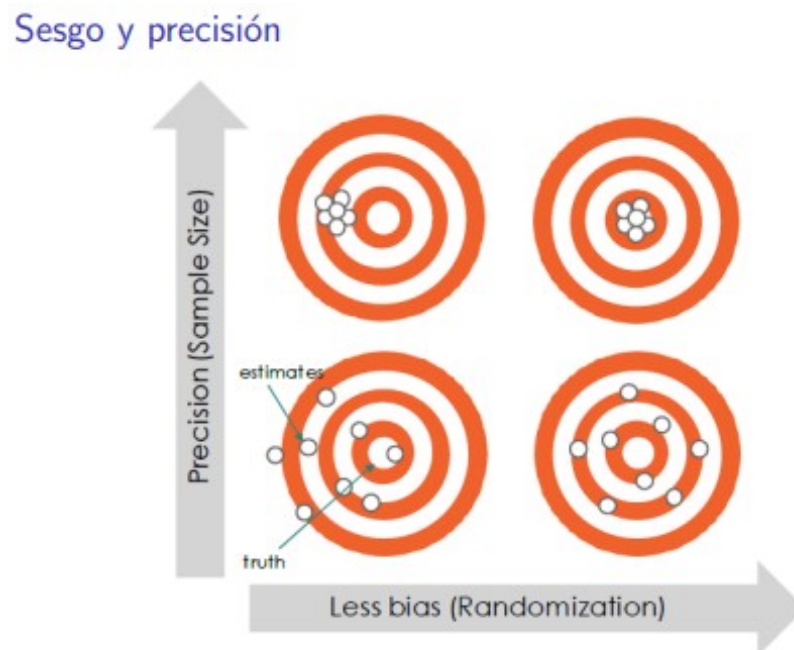


Figura 1: Relación entre precisión (arriba-abajo) y sesgo (izquierda-derecha).

Intuición:

- La precisión es como disparar muchas veces al mismo lugar; el sesgo es qué tan cerca estás del **centro real**.
- Un buen experimento es como un arquero que lanza flechas todas juntas y justo en el blanco: **consistente y correcto**.

Evaluamos el efecto de traer tutores a las escuelas

- Después de la prueba, se comparan **grupo control** y **grupo tratamiento**.
- La media del tratamiento resulta ser **6 puntos porcentuales más alta** que la media del grupo de control.

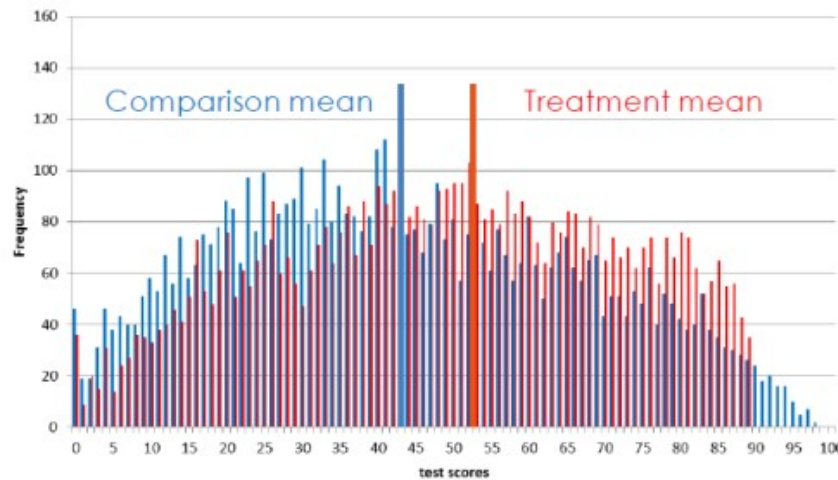


Figura 2: Distribución de puntajes: comparación entre media del grupo de control y del tratamiento.

- Pregunta central: ¿es este impacto **estadísticamente significativo**? ¿Sí, no, o no es posible determinarlo?

Intuición:

- Ver una media más alta no basta: necesitamos saber si la diferencia es **real** o solo fruto del azar.
- Es como lanzar una moneda muchas veces: un resultado con más “caras” puede parecer un patrón, pero solo la estadística nos dice si lo es en serio.

Diferencia entre las medias muestrales

- La diferencia entre la media del **grupo tratamiento** y la del **grupo control** se interpreta como el **efecto estimado**.
- En este caso, la diferencia corresponde a aproximadamente **6 puntos**.

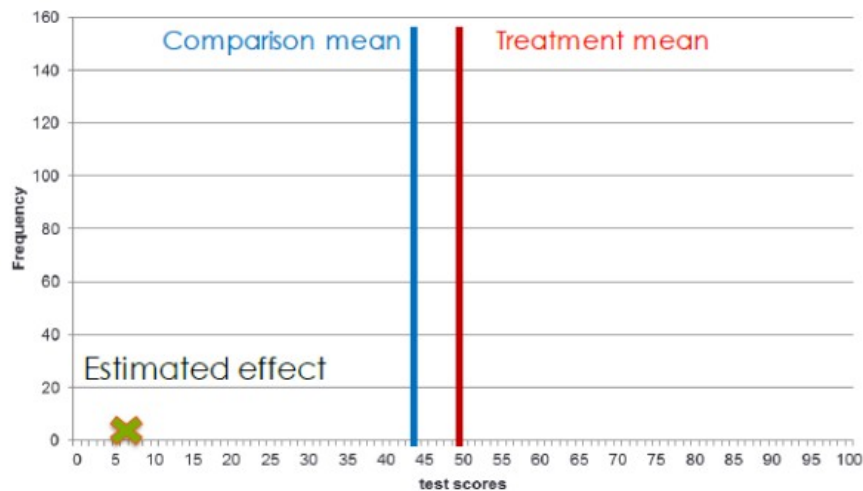


Figura 3: Comparación gráfica de medias muestrales: el efecto estimado es la distancia entre ambas líneas.

Intuición:

- La estimación es simplemente la **distancia** entre la media de los tratados y la de los controles.
- Es como comparar el promedio de altura entre dos equipos: la diferencia directa refleja el **efecto medio**.

¿Qué pasaría si corriéramos un segundo experimento?

- Si se repite el experimento con una nueva muestra, las **medias muestrales** de tratamiento y control pueden cambiar.
- El **efecto estimado** en este segundo experimento no será exactamente igual al del primero.

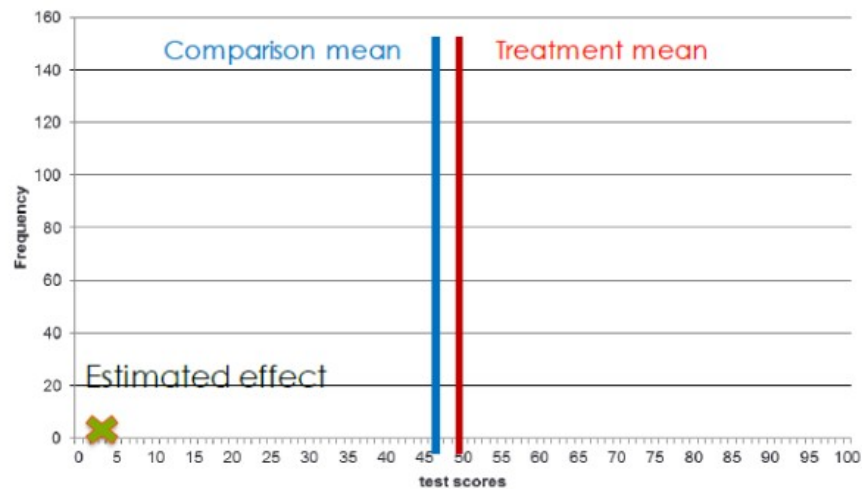


Figura 4: Un segundo experimento puede producir medias y efectos estimados diferentes debido al azar muestral.

Intuición:

- Cada experimento es como volver a **lanzar los dados**: los resultados pueden variar un poco aunque la verdad subyacente sea la misma.
- Esto refleja la **variabilidad muestral**: distintos grupos de estudiantes darán estimaciones ligeramente distintas.

Múltiples experimentos: una distribución de las estimaciones

- Si repitiéramos el experimento muchas veces, cada muestra daría un **efecto estimado** ligeramente distinto.
- Estos efectos se agrupan en una **distribución muestral**, que refleja la variabilidad inherente al azar.
- La forma de esta distribución permite evaluar la **precisión** de la estimación y construir intervalos de confianza.

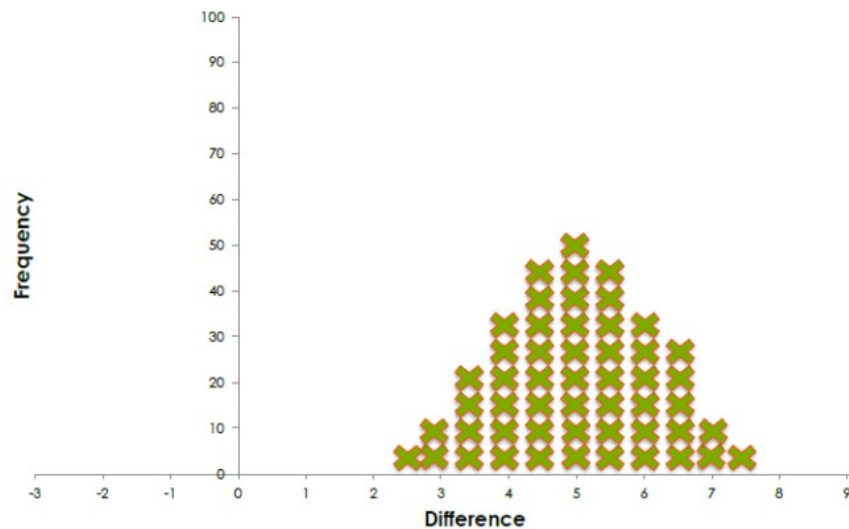


Figura 5: Distribución de los efectos estimados en múltiples repeticiones del experimento.

Intuición:

- Es como lanzar muchas veces un dado cargado: cada tirada cambia, pero la mayoría de los resultados se concentran cerca del valor real.
- Esta distribución nos dice qué tan **confiables** son los promedios que obtenemos y cuánto fluctúan por azar.

Distribución de las estimaciones si el verdadero efecto $= \beta$

- Los **efectos estimados** se concentran alrededor del verdadero parámetro β .
- Bajo ciertos supuestos, esta distribución sigue una forma **normal**, gracias al **Teorema Central del Límite**.
- Por eso, aunque cada experimento da un resultado distinto, en promedio los estimadores se agrupan cerca del valor real.

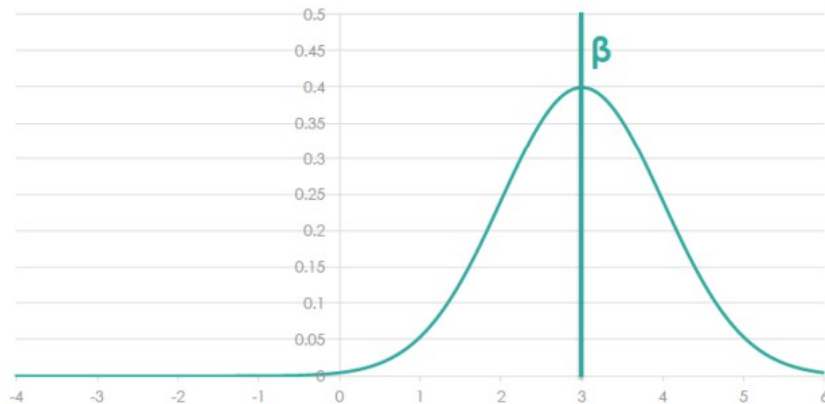


Figura 6: Distribución normal de los estimadores alrededor del efecto verdadero β .

Intuición:

- Es como lanzar una moneda muchas veces: la proporción de caras varía, pero se concentra alrededor de 0.5.
- Aquí, los estimadores fluctúan por azar, pero en promedio apuntan al **valor verdadero**.

Distribución de las estimaciones si el verdadero efecto = 0

- Cuando el **efecto verdadero es nulo**, los estimadores se concentran alrededor de 0.
- La dispersión de los resultados entre experimentos refleja el **error estándar**.
- En promedio, el estimador no está sesgado: su media coincide con el verdadero valor (0).

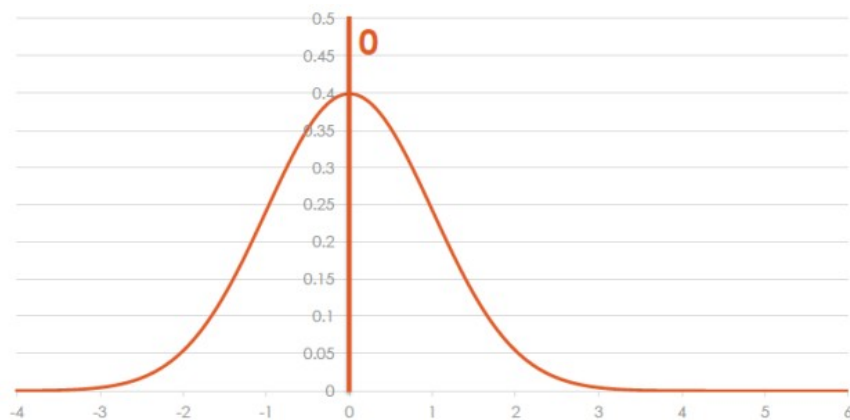


Figura 7: Distribución normal de los estimadores alrededor de 0 cuando el efecto verdadero es nulo.

Intuición:

- Como lanzar un dado balanceado: los resultados varían, pero en promedio no se inclinan hacia ningún lado.
- Aquí, cada experimento da un estimado distinto, pero el **promedio global es cero**.

Dos distribuciones bajo dos hipótesis distintas

- La **hipótesis nula** (H_0): el efecto verdadero es 0.
- La **hipótesis alternativa** (H_β): el efecto verdadero es distinto de 0 (por ejemplo, $\beta > 0$).
- Cada curva muestra la **distribución de los estimadores** si una u otra hipótesis fuera cierta.
- La superposición indica que un mismo valor estimado puede ser consistente con ambas hipótesis.

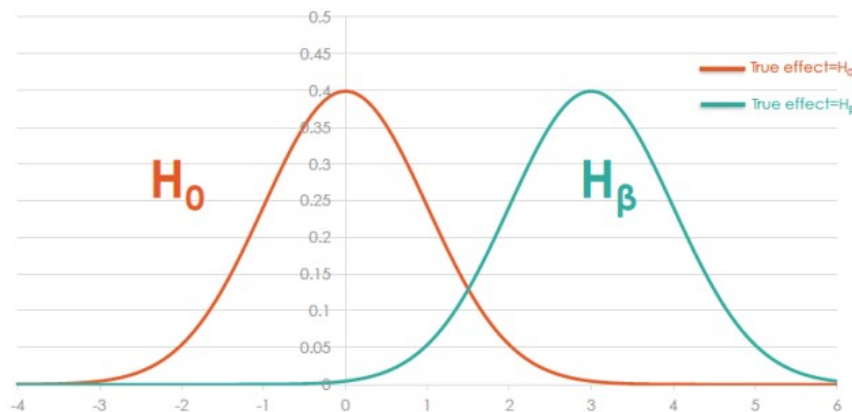


Figura 8: Distribuciones de los estimadores bajo H_0 (efecto nulo) y bajo H_β (efecto positivo).

Intuición:

- Como dos “mundos posibles”: en uno el efecto es 0, en el otro es positivo.
- Observar un estimado nos hace preguntarnos: ¿es más probable que provenga del mundo H_0 o del mundo H_β ?

Si corremos un solo experimento

- Al correr un solo experimento obtenemos una estimación puntual $\hat{\beta}$.
- La pregunta es: ¿cómo sabemos si este valor es estadísticamente significativo?
- Para responder, comparamos $\hat{\beta}$ con la distribución de los estimadores bajo H_0 .

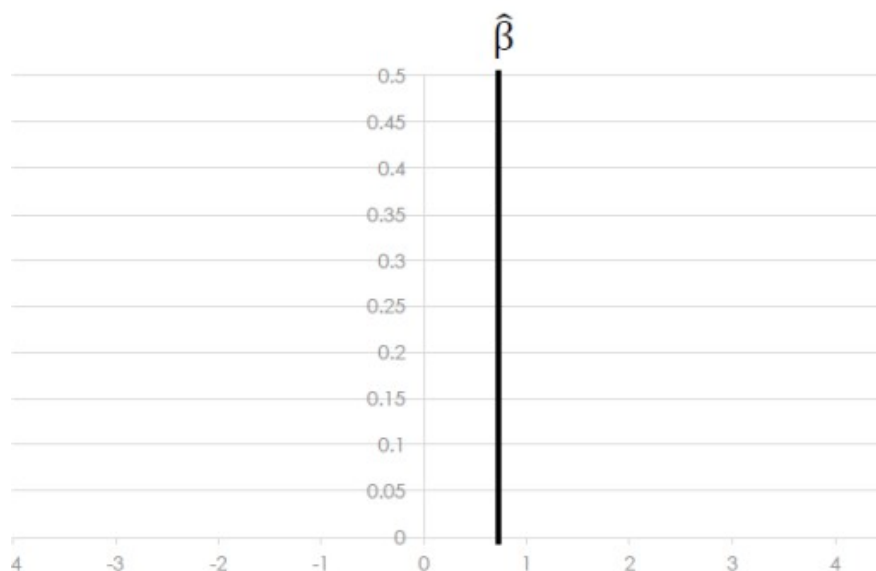


Figura 9: Una estimación puntual $\hat{\beta}$ a partir de un único experimento.

Intuición:

- Una sola flecha al blanco no nos dice si acertamos por suerte o por efecto real.
- Necesitamos comparar $\hat{\beta}$ con lo que esperaríamos si el efecto verdadero fuera cero (H_0).

¿Nuestra estimación provino de H_β o H_0 ?

- Una misma estimación $\hat{\beta}$ puede ser consistente tanto con H_0 como con H_β , pero con distinta probabilidad.
- El punto $P1$ muestra la densidad bajo H_0 ; el punto $P2$, bajo H_β .
- La pregunta central es: **¿qué hipótesis hace más probable la observación?**

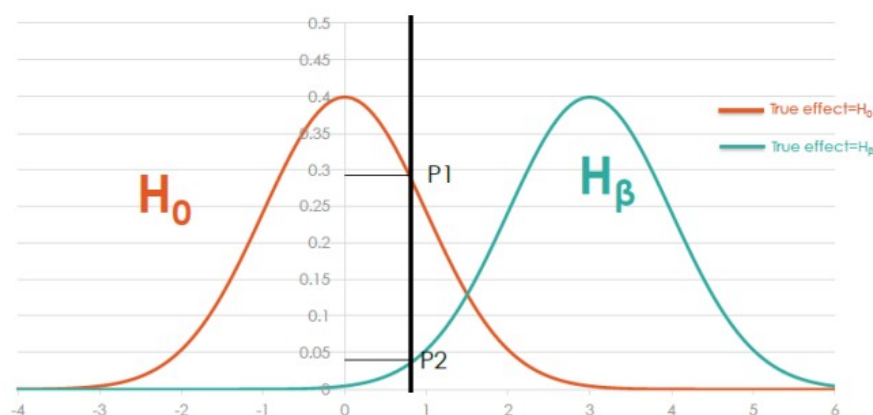


Figura 10: Comparación de probabilidades: densidad de la estimación bajo H_0 (P1) y bajo H_β (P2).

Intuición:

- Es como observar una ficha en la mesa: ¿salió de la bolsa roja (H_0) o de la bolsa azul (H_β)?
- No necesitamos certeza absoluta: lo clave es evaluar si podemos **descartar razonablemente** que provino de H_0 .

Nivel de significancia del 5 %

- Fijamos un umbral: $\alpha = 0,05$.
- Definimos dos **valores críticos** que delimitan la región donde no podemos rechazar H_0 .
- Cualquier estimación que caiga entre estas dos líneas críticas se interpreta como **no estadísticamente diferente de 0**.

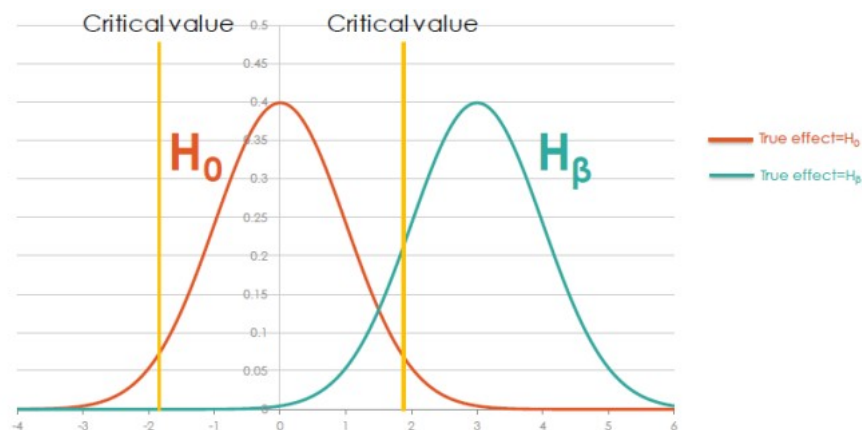


Figura 11: Regla de decisión con $\alpha = 0,05$: los valores entre las líneas críticas no permiten rechazar H_0 .

Intuición:

- Es como un margen de tolerancia: si la evidencia cae en esa “zona gris”, no podemos decir con confianza que el efecto es distinto de cero.
- Solo los valores fuera de esa franja se consideran suficientemente extremos para concluir que H_0 es improbable.

Valor crítico

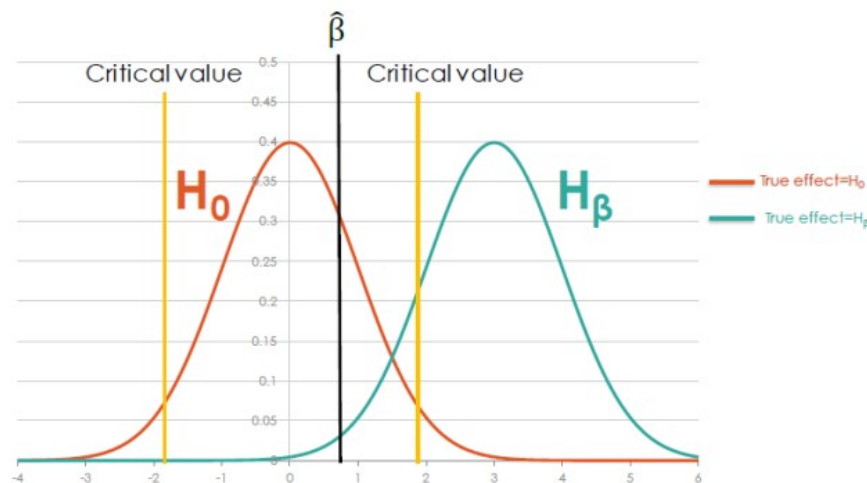
- **Definición:** El valor crítico es el tamaño del efecto estimado que corresponde exactamente con el nivel de significancia α .
- Sirve como punto de referencia para decidir si un resultado es suficientemente extremo como para rechazar H_0 .

- Si estamos probando:
 - El efecto es mayor que 0.
 - Y queremos significancia al nivel del 95 %.
 - Entonces, el **valor crítico** es la estimación donde exactamente 5 % del área bajo la curva se encuentra a la derecha.

Intuición:

- Es como fijar una “meta mínima”: solo los resultados más grandes que este umbral se consideran evidencia suficiente contra H_0 .
- Visualmente, el valor crítico es la frontera entre la región de aceptación y la de rechazo de H_0 .

¿Es β significativamente diferente de 0, al 5 % de significancia?



- La estimación $\hat{\beta}$ se compara con los valores críticos (líneas naranjas).
- Si $\hat{\beta}$ cae fuera del intervalo delimitado por los valores críticos, rechazamos H_0 .
- Si cae dentro, no podemos rechazar H_0 al 5 % de significancia.

Intuición:

- Los valores críticos son los “umbrales” de evidencia: solo los resultados suficientemente extremos permiten concluir que el efecto es distinto de 0.
- Aquí la pregunta clave es: *¿dónde cae nuestra estimación relativa a esos umbrales?*

Prueba de hipótesis

- En derecho penal, la mayoría de las instituciones siguen la regla de: **inocente hasta que se demuestre lo contrario**".
- La presunción es que el acusado es inocente y la carga está en el fiscal para demostrar la culpabilidad.
 - El jurado o juez inicia con la **hipótesis nula** de que el acusado es inocente.
 - El fiscal sostiene la hipótesis alternativa: que el acusado es culpable.

Intuición:

- La hipótesis nula funciona como la "base": se mantiene hasta que haya suficiente evidencia en contra.
- Igual que en un juicio, solo rechazamos H_0 (inocente) si la evidencia es muy fuerte a favor de H_1 (culpable).

Prueba de hipótesis (evaluación de programas)

- En evaluación de programas, en lugar de la "presunción de inocencia", la regla es: **presunción de cero efecto**".
- La **hipótesis nula** (H_0) es que el programa no tuvo impacto.
- La carga de la prueba es demostrar que sí hubo un impacto.
- **Excepción:** en programas con componentes múltiples, H_0 puede ser que un componente no cambie el efecto respecto a otro.
Ejemplo: un programa de *dinero en efectivo + capacitación* podría tener como H_0 que su efecto es igual al de un programa con *solo capacitación*.

Intuición:

- H_0 funciona como la "posición base" (no hay impacto).
 - Solo si la evidencia es suficientemente fuerte rechazamos H_0 y concluimos que el programa sí cambió algo.
-
- Si es muy improbable (menos del 5 % de probabilidad) que la diferencia entre grupo control y tratamiento se deba exclusivamente a la casualidad:
 - Rechazamos la hipótesis nula.
 - Es decir, concluimos que el programa tiene un impacto **estadísticamente significativo**.

- Importante: **estadísticamente significativo** \neq **más probable**.
- Puede ocurrir que sea más probable que el programa funcione a que no lo haga, pero aún así no se considere **estadísticamente significativo**.

Intuición:

- La significancia estadística es una regla formal: no basta con que algo “parezca más probable”.
- Sirve para reducir errores y exigir evidencia fuerte antes de rechazar la hipótesis nula.

¿Qué es el nivel de significancia?

- **Error Tipo I:** rechazar la hipótesis nula a pesar de que es verdadera.
- **Nivel de significancia:** probabilidad de rechazar la hipótesis nula cuando en realidad es verdadera.
- Tradicionalmente se fija en **5 %**.
- Esto implica aceptar un 5 % de probabilidad de cometer un Error Tipo I.
- Es decir, el 5 % del tiempo diremos que un programa tuvo impacto, aunque en realidad no lo tuvo.

Intuición:

- El nivel de significancia es como el *umbral de confianza* que ponemos antes de afirmar que algo funciona.
- Aceptamos que, a veces, podemos equivocarnos —pero solo en una proporción pequeña (5 %).

¿Qué es poder estadístico?

- **Error Tipo II:** no rechazar la hipótesis nula (concluir que no hay diferencia entre control y tratados) cuando en realidad es falsa.
- **Poder:** probabilidad de detectar un efecto real (rechazar H_0 cuando H_0 es falsa).
- Tradicionalmente, se busca un **80 %** de poder estadístico (otros apuntan a **90 %**).
- Un nivel bajo de poder significa que podríamos no encontrar un efecto significativo, aun cuando sí existe.

Intuición:

- El poder estadístico mide la *capacidad del experimento para no perderse un efecto real*.
- Como un radar: si el radar es débil (bajo poder), puede que no detecte un avión aunque esté allí.

Cuatro resultados de las pruebas de hipótesis

- Cuando el tratamiento es **efectivo** (H_0 falsa):
 - Si la prueba es **significativa** (rechaza H_0): obtenemos un resultado correcto con probabilidad $1 - \kappa$ (**poder**).
 - Si la prueba es **no significativa** (no se rechaza H_0): incurrimos en un **Error Tipo II** con probabilidad κ .
- Cuando el tratamiento **no es efectivo** (H_0 verdadera):
 - Si la prueba es **significativa** (rechaza H_0): incurrimos en un **Error Tipo I** con probabilidad α .
 - Si la prueba es **no significativa** (no se rechaza H_0): obtenemos un resultado correcto con probabilidad $1 - \alpha$.

		La verdad:	
		El tratamiento es efectivo (H_0 falsa)	El tratamiento no es efectivo (H_0 verdadera)
Prueba estadística	Significativo (Rechaza H_0)	 Probabilidad= $1-\kappa$ (poder)	Error Tipo I Probabilidad= α 
	No significativo (No se rechaza H_0)	Error Tipo II Probabilidad= κ 	 Probabilidad= $(1-\alpha)$

Figura 12: Resumen de los posibles resultados de una prueba de hipótesis.

Intuición:

- Piensa en un detector de humo: a veces suena aunque no haya fuego (**Error Tipo I**), o no suena aunque sí haya fuego (**Error Tipo II**).
- El poder estadístico refleja qué tan bueno es el detector para sonar cuando realmente hay fuego (detectar un efecto real).

¿Cuántas veces se rechazaría la hipótesis nula, si es que H_β es cierta?

- El área sombreada representa el **poder estadístico**: la fracción de veces que encontramos que H_β es distinto de 0 cuando en verdad el efecto es β .

- El poder es, por tanto, la **probabilidad de rechazar H_0 cuando esta es falsa**.
- La línea amarilla marca el **valor crítico** a partir del cual concluimos que el efecto es significativo.

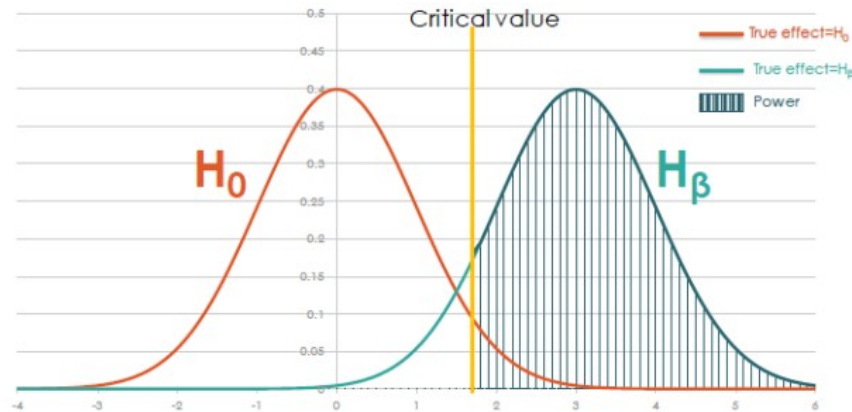


Figura 13: El área bajo H_β más allá del valor crítico indica el poder estadístico.

Intuición:

- Es como un examen de antidoping: el poder mide cuántas veces detectamos correctamente a alguien que sí consumió (efecto real).
- Un poder alto significa que rara vez dejamos pasar por “negativo” a alguien que en verdad era positivo.

¿Cuáles son los determinantes del poder estadístico?

- Los **determinantes del poder** son los factores que modifican la proporción de hipótesis de investigación que quedan en la zona sombreada, es decir, la parte a la derecha (o izquierda) de la curva de H_0 .
- Comprender estos determinantes es clave para **diseñar experimentos más potentes** y con mayor capacidad de detectar efectos reales.

Intuición:

- El poder depende de qué tan fácil es distinguir la señal del ruido: más muestra, menor ruido y más chances de ver el efecto.
- Es como usar una cámara: con mayor resolución (poder estadístico), captas detalles que con una cámara borrosa pasarían desapercibidos.

Poder estadístico: principales determinantes

- Mayor **superposición de las curvas** implica menor poder estadístico.
- Los factores que determinan el grado de superposición entre H_0 y H_β son:
 1. **Tamaño del efecto**: efectos más grandes son más fáciles de detectar.
 2. **Tamaño de la muestra**: muestras mayores reducen la incertidumbre.
 3. **Varianza**: menor dispersión de los datos aumenta el poder.
 4. **Proporción de la muestra en T vs. C**: la asignación balanceada maximiza eficiencia.
 5. **Clustering (clusterización)**: aumenta la correlación interna y reduce poder.

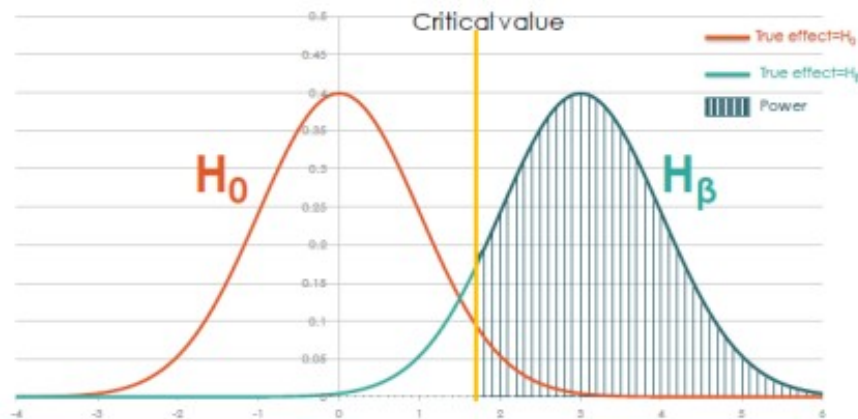


Figura 14: Factores que determinan el solapamiento de las distribuciones y, por tanto, el poder.

Intuición:

- Más datos y menos ruido hacen que la señal se vea más clara: como subir el volumen de una canción en medio del tráfico.
- Si las curvas se separan más (efecto grande, muestra grande), es mucho más difícil confundir H_0 con H_β .

Tamaño del efecto = 1 × error estándar

- El **tamaño del efecto** puede medirse en múltiplos del **error estándar**.
- Si el efecto verdadero es igual a 1 error estándar, las distribuciones de H_0 y H_β se **solapan bastante**.
- Esto implica que distinguir entre hipótesis se vuelve más difícil, reduciendo el poder.

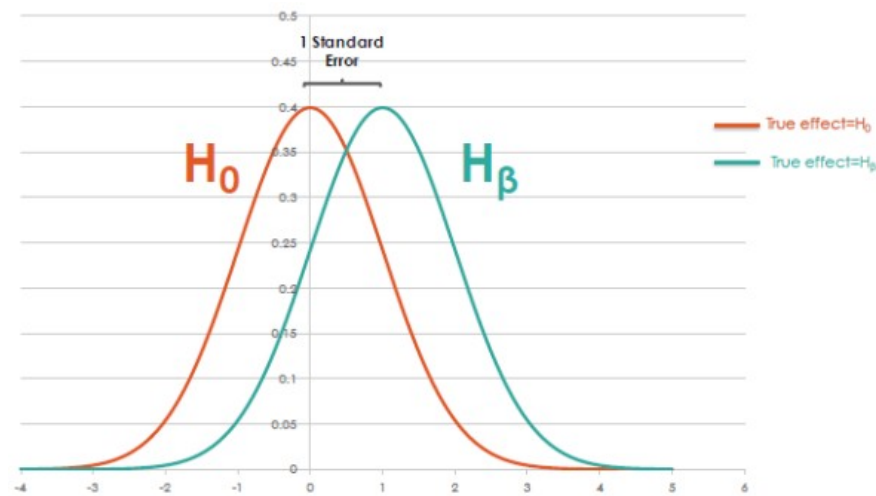


Figura 15: Cuando el efecto es de 1 error estándar, la superposición entre H_0 y H_β es significativa.

Intuición:

- Imagínalo como dos grupos de personas con alturas distintas, pero que difieren en solo **unos centímetros**: cuesta más diferenciarlos.
- Cuanto más pequeño sea el efecto en relación al ruido (error estándar), más se confunden las distribuciones y menor el poder.

Tamaño del efecto = $1 \times$ error estándar

- Si el efecto verdadero equivale a **un error estándar**, la zona sombreada bajo H_β muestra la fracción de casos donde se rechaza H_0 .
- El **poder estadístico** es relativamente bajo, porque gran parte de las distribuciones de H_0 y H_β siguen solapadas.
- Esto refleja que con un efecto pequeño, el test solo detecta diferencias en una porción limitada de escenarios.

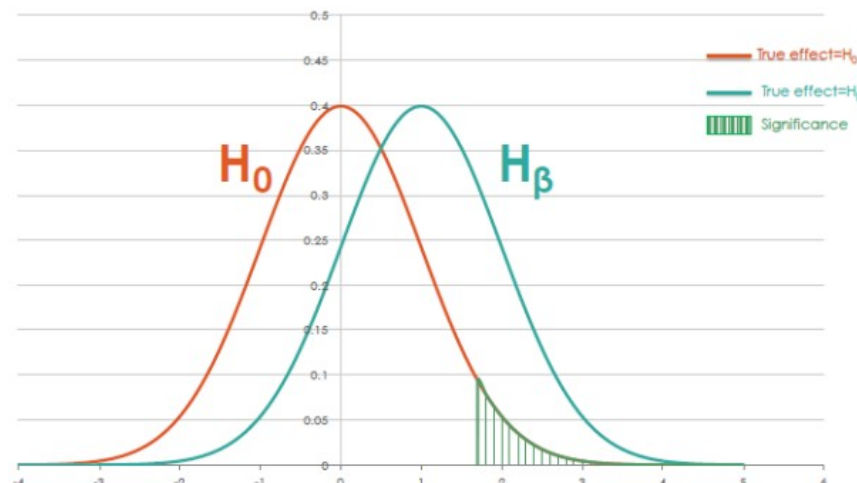


Figura 16: El área verde representa la probabilidad de rechazar H_0 cuando el verdadero efecto es H_β .

Intuición:

- Es como tratar de escuchar una canción suave en medio de mucho **ruido**: a veces la distingues, pero muchas veces no.
- Un efecto pequeño en relación al error estándar significa que el experimento tiene **bajo poder**.

Poder igual a 26 % si el verdadero impacto es $1 \times$ error estándar

- Con un **efecto verdadero de 1 error estándar**, el poder estadístico resulta ser apenas **26 %**.
- Esto implica que en el **74 % de los casos**, no se rechazaría la hipótesis nula aunque el efecto exista.
- La figura muestra el área sombreada bajo H_β , correspondiente a la probabilidad de detectar el efecto real.

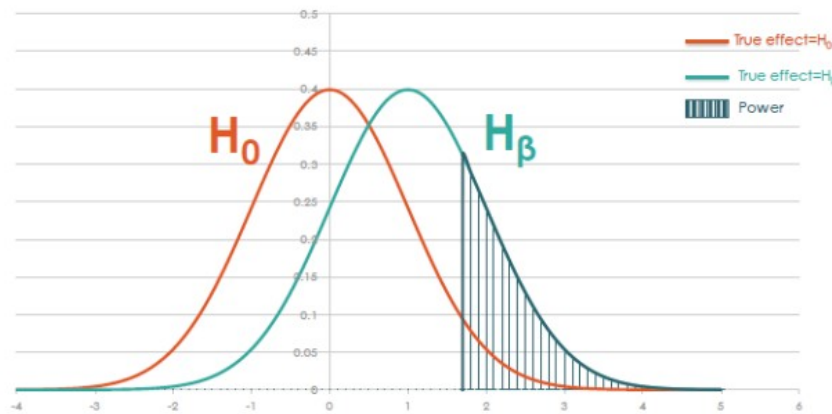


Figura 17: Con un efecto de 1 error estándar, la hipótesis nula solo se rechaza en el 26 % de los experimentos.

Intuición:

- Es como lanzar una moneda cargada pero muy poco: la mayoría de las veces sigue pareciendo una moneda justa.
- Un poder de 26 % significa que tu **experimento está mal calibrado**: la mayor parte del tiempo no detecta un efecto que sí existe.

Tamaño del efecto = $3 \times$ error estándar. Poder = 91 %

- Con un **efecto de 3 errores estándar**, las distribuciones de H_0 y H_β están claramente separadas.
- El **poder estadístico** alcanza el 91 %, lo que significa que casi siempre se detecta el efecto verdadero.
- A mayor **tamaño del efecto**, menor solapamiento entre distribuciones y, por tanto, mayor precisión en el test.

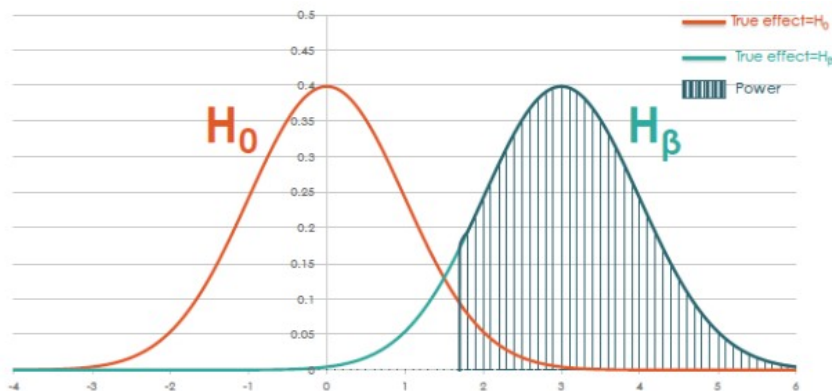


Figura 18: Un efecto tres veces mayor que el error estándar produce un poder muy alto (91 %).

Intuición:

- Es como comparar a una persona adulta con un niño: la diferencia es tan clara que casi nunca te confundes.
- Con un **efecto grande**, el experimento funciona de manera mucho más confiable y detecta el impacto verdadero casi siempre.

Tamaño del efecto y adopción (take-up)

- Supongamos que el impacto esperado en los participantes es de **3 errores estándar**.
- Pero si solo un **tercio de los individuos** adopta o participa, el efecto observado se reduce a **1/3 del tamaño**.
- Esto implica que el poder estadístico vuelve a caer drásticamente, en este caso a **26 %**.

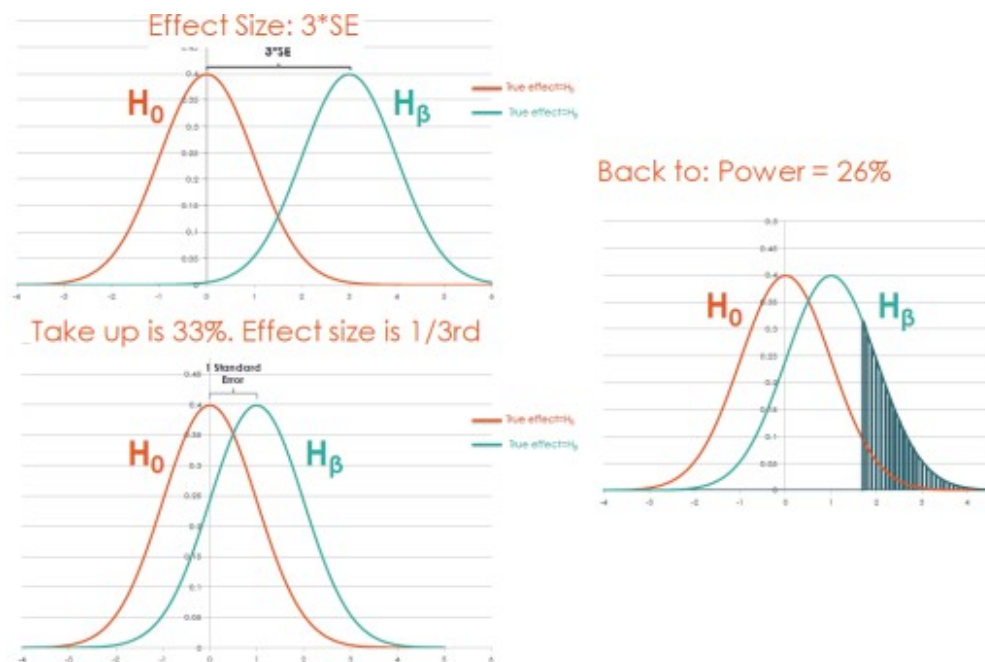


Figura 19: Con adopción parcial (33 %), el efecto efectivo se reduce y el poder baja nuevamente al 26 %.

Intuición:

- Es como tener un remedio muy potente, pero que solo toma una parte de los pacientes: el **efecto promedio se diluye**.
- Aun si el impacto individual es grande, la baja participación hace que el experimento pierda fuerza.

Al incrementar el tamaño de la muestra...

- (a) Disminuye el sesgo
- (b) Incrementa precisión
- (c) Ambas
- (d) Ninguna
- (e) No sé

Al incrementar el tamaño de la muestra...

- (a) Las curvas se separan
- (b) Las curvas se acercan
- (c) Las curvas se ensanchan
- (d) Las curvas se estrechan
- (e) No sé

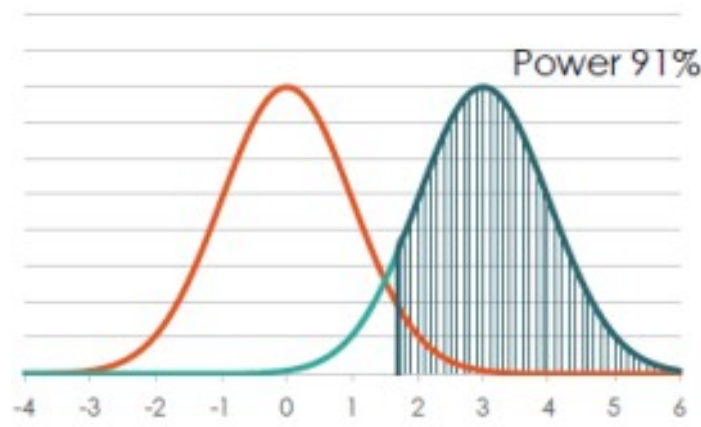


Figura 20: Un mayor tamaño de muestra estrecha las curvas, aumentando la precisión y el poder (ej. 91 %).

Intuición:

- Más observaciones = menos ruido: las distribuciones se **estrechan** porque reducimos la variabilidad.
- Es como medir la temperatura varias veces: al aumentar las mediciones, el promedio se vuelve **más preciso**.

Tamaño de la muestra

- Con un **tamaño de efecto = 1 error estándar** y una muestra de **4000**, el poder es de apenas **64 %**.
- Al aumentar la muestra a **9000**, el poder crece hasta **91 %**.
- Esto muestra que una mayor muestra estrecha las distribuciones y mejora la probabilidad de rechazar H_0 cuando corresponde.

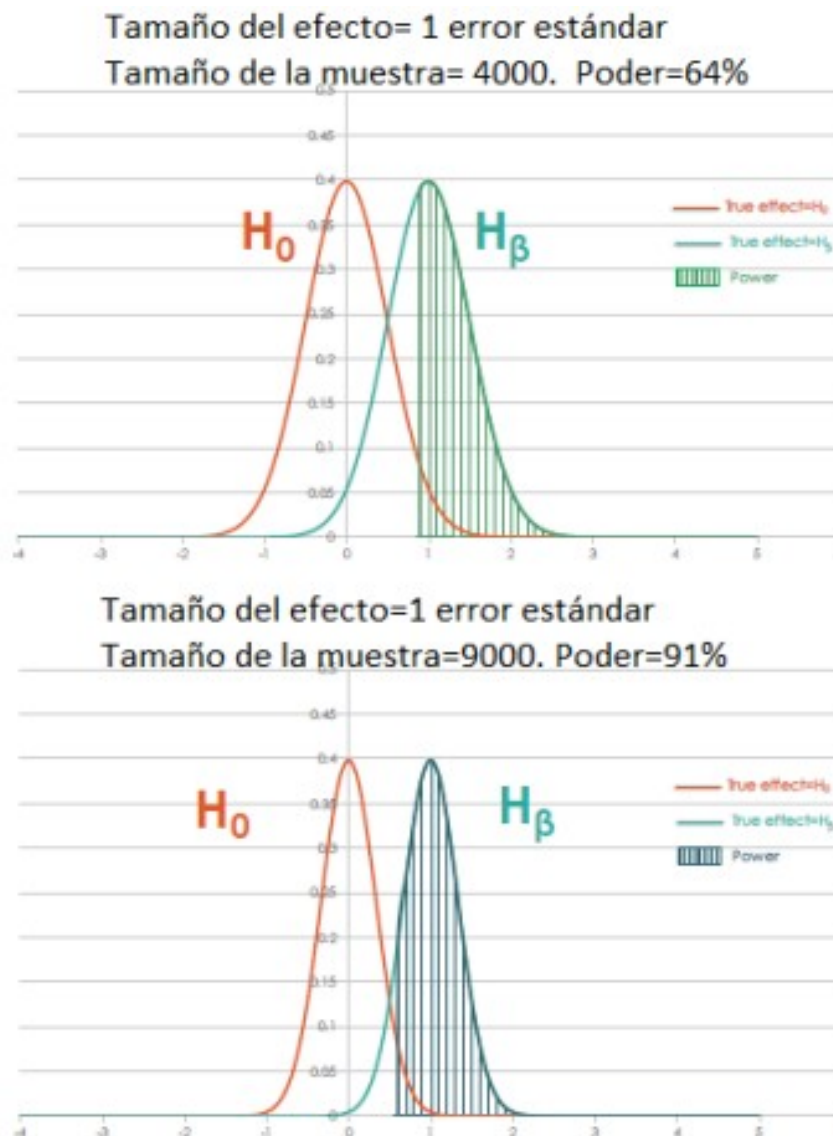


Figura 21: A mayor tamaño de muestra, las distribuciones se estrechan y el poder estadístico aumenta (64 % \rightarrow 91 %).

Intuición:

- Con pocas observaciones, el “ruido” hace difícil distinguir si hay un efecto.
- Una muestra grande es como tener una foto en alta resolución: todo se ve con mayor claridad y precisión.

¿Cómo un incremento en la varianza de la población afecta las curvas de distribución de las estimaciones?

- (a) Las curvas se separan
- (b) Las curvas se acercan
- (c) Las curvas se ensanchan
- (d) Las curvas se estrechan
- (e) No sé

Intuición:

- Mayor **varianza** significa más dispersión: las distribuciones se hacen más anchas y con colas más largas.
- Es como medir con una regla torcida: el promedio puede estar bien, pero los datos quedan mucho más **esparcidos**.

Varianza

- Estimaciones más agrupadas → mayor poder estadístico.
- Muestra con poca varianza: distribuciones estrechas, fácil distinguir H_0 de H_β .
- Muestra con mucha varianza: distribuciones más dispersas, mayor solapamiento y menor capacidad de detección.
- Estimaciones más dispersas → menor poder.



Figura 22: Comparación entre baja varianza (arriba) y alta varianza (abajo): más dispersión implica menor poder estadístico.

Intuición:

- Con **poca varianza**, es como ver dos colores bien definidos: la diferencia salta a la vista.
- Con **mucha varianza**, los colores se mezclan y cuesta saber cuál es cuál: el test pierde fuerza.

Varianza y poder: intuición

- Un programa busca aumentar la altura de los niños.
- Existe una gran cantidad de variación en la altura de la población.
- Al final del período, los niños en el grupo de tratamiento son más altos que en el grupo de control.
- ¿Será porque comenzamos con niños más altos o porque el programa funcionó?
- Necesitamos una muestra grande para poder averiguar esto.

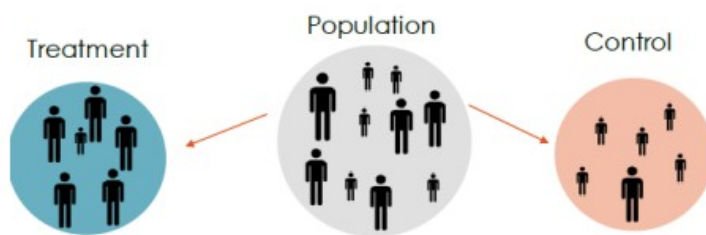


Figura 23: Ejemplo intuitivo: variabilidad inicial puede confundir el efecto real del tratamiento.

Intuición:

- Con mucha varianza, es difícil separar el efecto del programa de las diferencias iniciales.
 - Una **muestra grande** ayuda a “promediar” la variación y revelar si el tratamiento realmente funcionó.
-
- Si todos los miembros de la población subyacente son similares en estatura al inicio del programa, sería fácil averiguar el efecto.
 - Necesitaríamos una muestra más pequeña (o, con el mismo tamaño de muestra, tendríamos un mayor poder estadístico).
 - Sería más fácil concluir que las variaciones observadas entre el grupo tratado y el de control son resultado del programa.

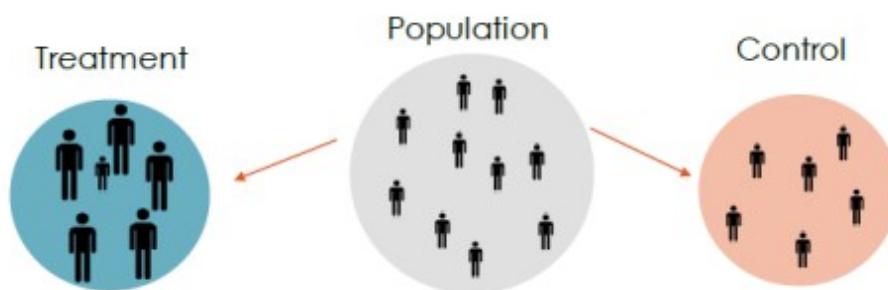


Figura 24: Menor varianza inicial implica más claridad en la detección del efecto.

Intuición:

- Si la población es homogénea, el “ruido” es bajo y se necesita menos muestra para identificar el impacto.
- Menor varianza = más fácil distinguir efecto verdadero del programa.

Proporción de la muestra en T vs. C

- División de la muestra: 50 % C, 50 % T. Poder = 91 %.
- La igualdad en la proporción da distribuciones con la misma “anchura”.

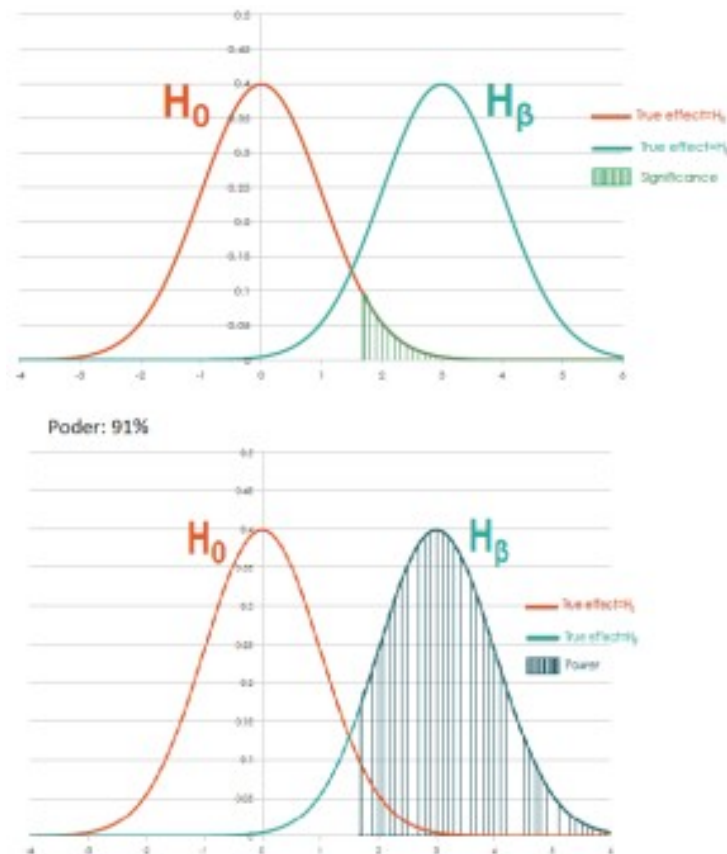


Figura 25: Poder estadístico con división equilibrada de la muestra entre control y tratamiento.

Intuición:

- Una división balanceada entre tratamiento y control maximiza la precisión de la comparación.
- Si la muestra está desbalanceada, aumenta la varianza y disminuye el poder.

¿Qué pasa si la división de la muestra no es del 50:50?

- Por ejemplo, si es de 25 % C y 75 % T, el poder baja a 83 %.
- Distribuciones desiguales no son eficientes → menor poder.

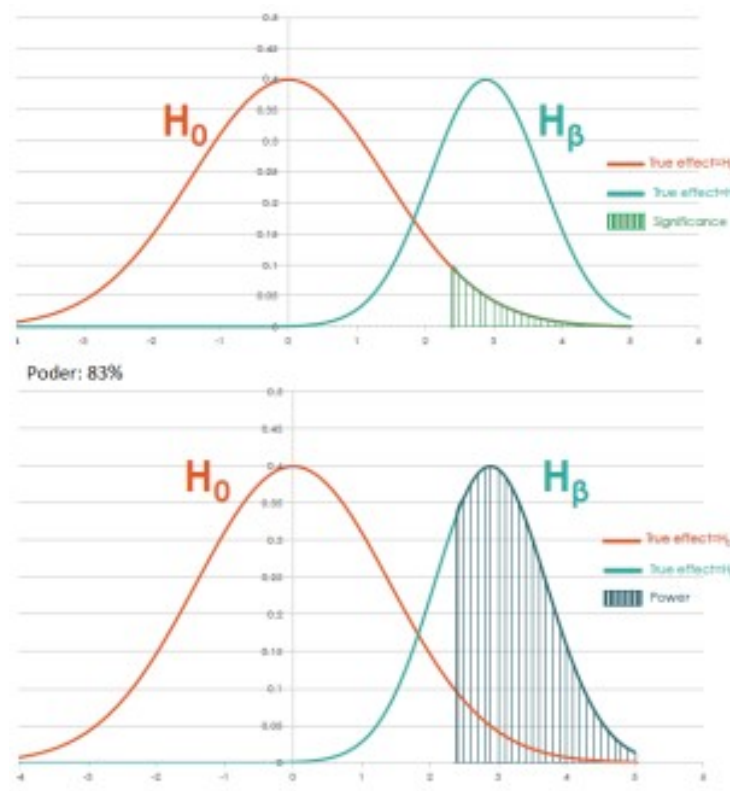


Figura 26: Poder estadístico disminuye cuando la división no es balanceada entre grupos.

Intuición:

- Un grupo pequeño (ej. 25 % en control) aporta menos información para la comparación.
- Esto aumenta la varianza y reduce la probabilidad de detectar un efecto real.

¿Qué tanto es muy desequilibrado?

Bloom (2006): "Debido a que la precisión se erosiona lentamente hasta que el grado de desequilibrio se vuelve extremo (o más o menos $p \leq 0,2$ o $p \geq 0,8$), hay suficiente margen de maniobra para el uso de una asignación desequilibrada".

- Cuestiones políticas pueden dictar un grupo pequeño de control.
- Los costos pueden dictar un grupo pequeño de tratamiento.

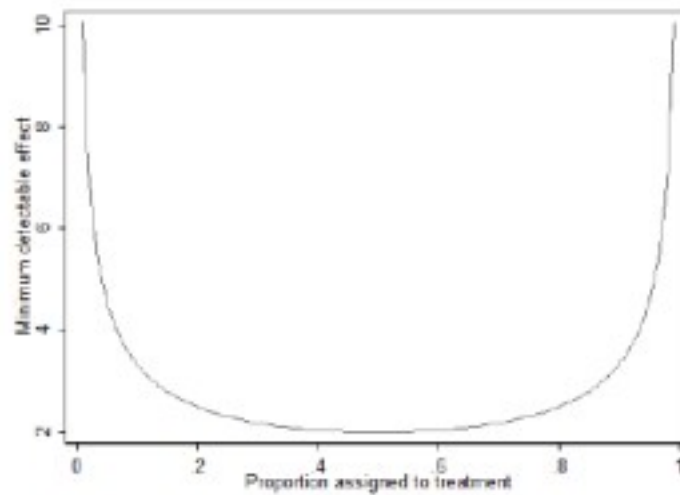


Figura 27: Relación entre proporción asignada y varianza del estimador.

- Sin embargo, dejando atrás estas consideraciones, para un tamaño de muestra dado, el poder se maximiza cuando la mitad de la muestra es asignada al tratamiento.

Ecuación de poder. Efecto Mínimo Detectable

$$EMD = \underbrace{t_{(1-\kappa)}}_{\text{Poder}} + \underbrace{t_{\alpha}}_{\text{Nivel de significancia}} \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{N}}$$

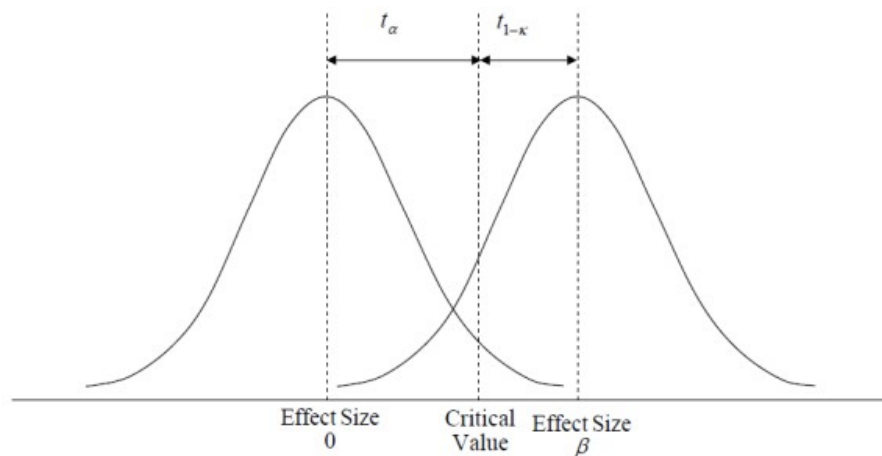


Figura 28: Representación gráfica del Efecto Mínimo Detectable (EMD).

- El **EMD** es el tamaño de efecto más pequeño que puede ser detectado con un nivel de confianza y poder dados.
- Depende de la **varianza** (σ^2), el **tamaño de la muestra** (N) y la **proporción de asignación** (P).

- Valores más altos de N o menor varianza reducen el EMD, permitiendo detectar efectos más pequeños.

Diseño agrupado. Definición

En muestreo:

- Cuando grupos de individuos (por ejemplo, escuelas, comunidades, etc.) son seleccionados al azar de la población, antes de seleccionar individuos para observación.

En evaluación aleatoria:

- Cuando grupos de individuos son asignados al azar a diferentes grupos de tratamiento.

Queremos saber qué tan cerradas serán las próximas elecciones nacionales

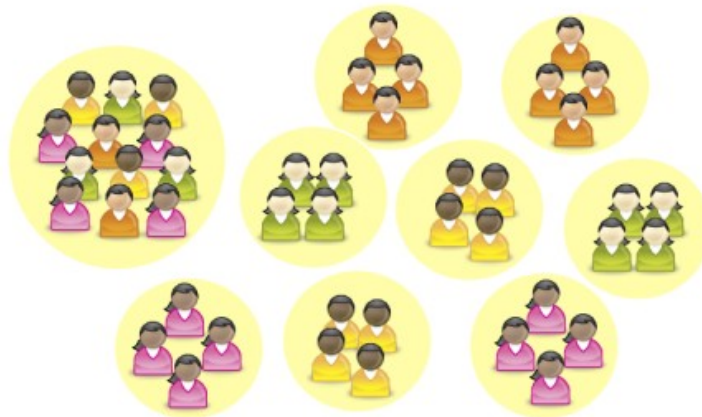
- **Método 1:** Seleccionar aleatoriamente 50 personas de toda la población
- **Método 2:** Seleccionar al azar 5 familias y pedir a 10 miembros de cada familia su opinión

Baja correlación dentro de la agrupación o clúster (ICC, ρ)



Intra-cluster correlation (ICC): mide qué tan parecidos son los individuos dentro de un mismo clúster. Si la correlación ρ es baja, cada observación aporta más información independiente.

- **Explicación:** Cuando ρ es cercano a 0, los individuos dentro del clúster se comportan como observaciones casi independientes.
- **Intuición:** Un grupo de personas con opiniones muy diversas equivale a tener más información útil, aunque estén en el mismo clúster.



Si la correlación intra-clúster ρ es alta, los individuos dentro del mismo grupo se parecen demasiado. Esto reduce la información efectiva, ya que cada observación adicional aporta poco valor nuevo.

Alta correlación dentro de la agrupación o clúster (ρ)

- **Explicación:** Con ρ alto, los individuos de un clúster se comportan casi igual, como si fueran una sola observación.
- **Intuición:** Es como encuestar a una familia donde todos responden lo mismo: aunque tengas muchas respuestas, la información es redundante.

¿Cómo ρ (ICC) influye en el poder estadístico?

- Para un tamaño de muestra determinado N , tenemos **menos poder** cuando randomizamos a nivel clúster (a menos que $ICC = 0$).
- Lo clave para determinar el poder es el **número de clústeres**, no el número de personas dentro de cada clúster.
- Fórmula de poder con diseño agrupado:

$$\frac{EMD}{\sqrt{1 + \rho(m - 1)}} = (t_{(1-\kappa)} + t_{\alpha}) \times \sqrt{\frac{1}{P(1 - P)}} \times \sqrt{\frac{\sigma^2}{N}}$$

- m : Tamaño promedio del clúster
- ρ : Correlación intra-clúster (ICC)
- **Explicación:** Un ρ alto inflará el denominador y reducirá el poder, ya que cada observación adicional dentro del clúster aporta información redundante.
- **Intuición:** Es como entrevistar a muchos hermanos de la misma familia: si todos responden parecido, no se gana información adicional útil.

Calcular poder con Stata

- Stata tiene un nuevo comando de **power**, donde señalas el tamaño de la muestra y te muestra el poder.
- **Limitación:** no permite el diseño agrupado.
- La mayoría todavía utiliza **sampsi**, **sampclus** (add-ons) o **clustersampsi**.
 - Por defecto calcula un poder del 90 % y 5 % de significancia con asignación equitativa.
- Ejemplo: para detectar un aumento en las puntuaciones promedio de las pruebas de 43 % a 45 % con un poder de 80 %:

```
sampsi 0.43 0.45, power(0.8) sd(0.05)
```

- **Explicación:** El comando estima automáticamente el tamaño de muestra requerido para alcanzar el poder deseado dado un efecto esperado.
- **Intuición:** Es como preguntarle a Stata: “¿Con cuántos alumnos necesito trabajar para estar 80 % seguro de detectar una mejora de 43 % a 45 % en los puntajes?”.