

Inferencia

2025-04-23

Intro

Para esa clase, y para ensayos aleatorios en general, recomiendo el paper **Using Randomization in Development Economics Research: A Toolkit** *Esther Duflo, Rachel Glennerster and Michael Kremer*

Prueba de Hipótesis Clásica

Prueba de Hipótesis Clásica

Empezamos con una hipótesis sobre β a probar, $H_0: \beta = \mu$

- Comúnmente, $H_0: \beta = 0$.

Elegimos una estadística de prueba.

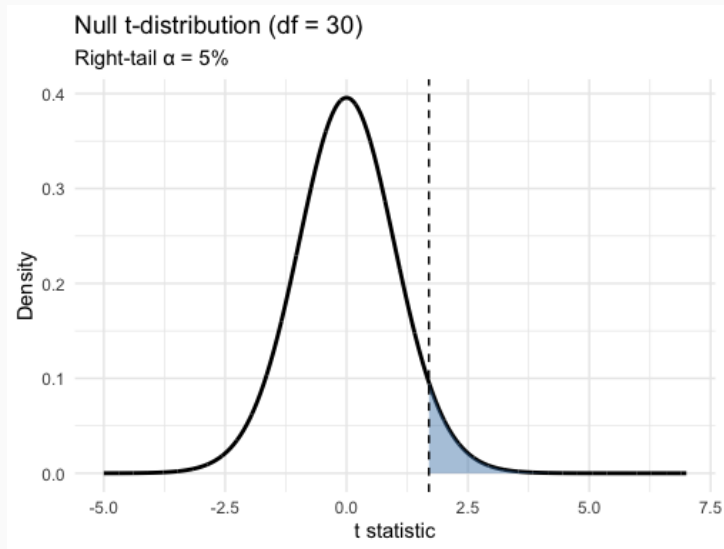
- Por ejemplo: $t = \frac{\hat{\beta} - \mu}{se(\hat{\beta})}$

Y fijamos el nivel de significancia deseado ($\alpha = 5\%$)

Podemos reconstruir la distribución de probabilidades de una estadística bajo ciertas hipótesis:

- H_0 es verdadera (o sea, $\beta = \mu$)
- N es grande (aproximación asintótica)

Prueba de Hipótesis Clásica



Rechazamos H_0 si la estadística t es muy extrema en comparación con valores más típicos si H_0 fuera verdad.

Bajo la hipótesis nula, la probabilidad de rechazar la nula es 5% (o α).

Si cambiamos la elección de *alpha* de 5% para 10%, que resulta?

- A: Tenemos menos probabilidad de rechazar una H_0 falsa
- B: Tenemos más probabilidad de rechazar una H_0 verdadera
- C: La potencia de la prueba aumenta
- D: B y C son verdad

La **potencia** de una prueba es la probabilidad de rechazar una hipótesis nula que es **falsa**.

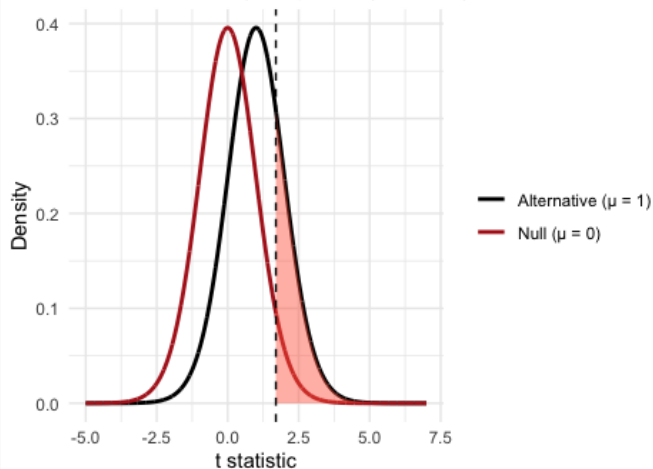
Una buena prueba tiene alta potencia: puede rechazar hipótesis nulas falsas con alta probabilidad.

La potencia depende de cual es el valor real de β , que no sabemos. Portanto, solo podemos calcular bajo la hipótesis de que β tiene un cierto valor.

Potencia de Prueba

Statistical Power Illustration

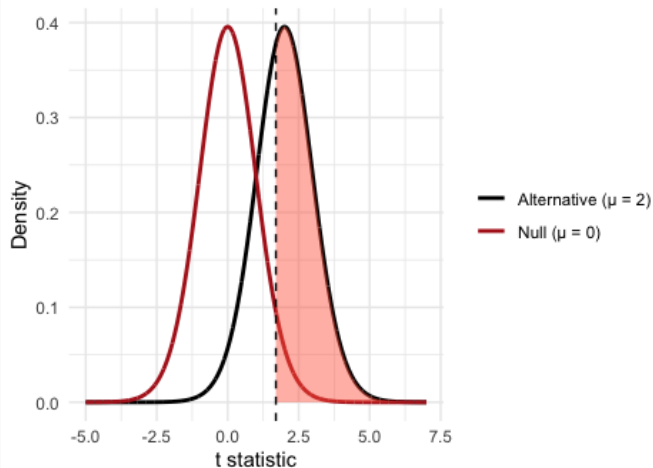
Red area = power ($\Pr[\text{reject } H_0 \mid H_1 \text{ true}]$)



Potencia de Prueba

Statistical Power Illustration

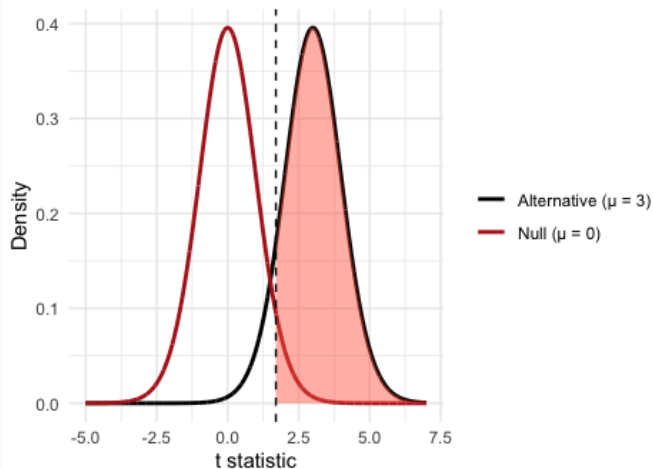
Red area = power ($\Pr[\text{reject } H_0 \mid H_1 \text{ true}]$)



Potencia de Prueba

Statistical Power Illustration

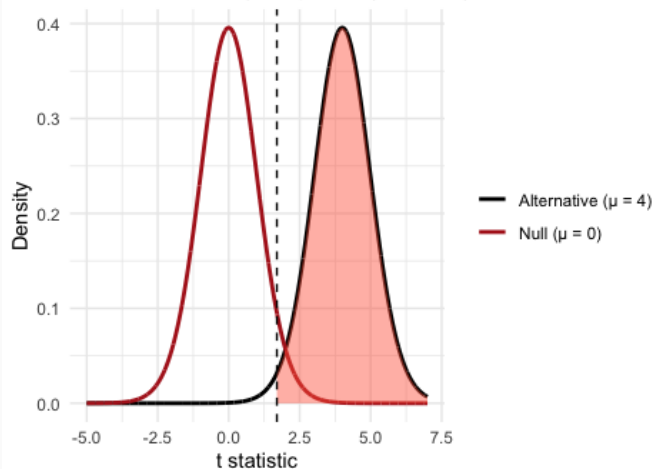
Red area = power ($\Pr[\text{reject } H_0 \mid H_1 \text{ true}]$)



Potencia de Prueba

Statistical Power Illustration

Red area = power ($\Pr[\text{reject } H_0 \mid H_1 \text{ true}]$)



Podemos calcular la potencia para diferentes valores y exprimirlos como una *curva de potencia*, o sea, la potencia como función del parámetro.

Cuales son verdad de una curva de potencia $CP(x)$, con $H_0: \beta = \mu_0$?

- A) $CP(\mu_0) = \alpha$
- B) $\lim_{x \rightarrow \infty} CP(x) = 1$
- C) $CP(x)$ es una función no decreciente de x
- D) $CP(x) < 1$ para todo x
- E) Para $x < \mu_0$, $CP(x) < \alpha$

Prueba de Hipótesis en Ensayos Aleatorios

Cuando trabajamos con ensayos aleatorios, podemos utilizar las mismas herramientas de prueba de hipótesis.

Pero, como podemos crear el diseño del ensayo, debemos pensar un poco sobre su potencia.

Vamos pensar en un ensayo aleatorio simple, donde se estima la regresión:

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

Vamos asumir que una proporción P de la muestra es tratada, y que la varianza de Y_i es σ^2 .

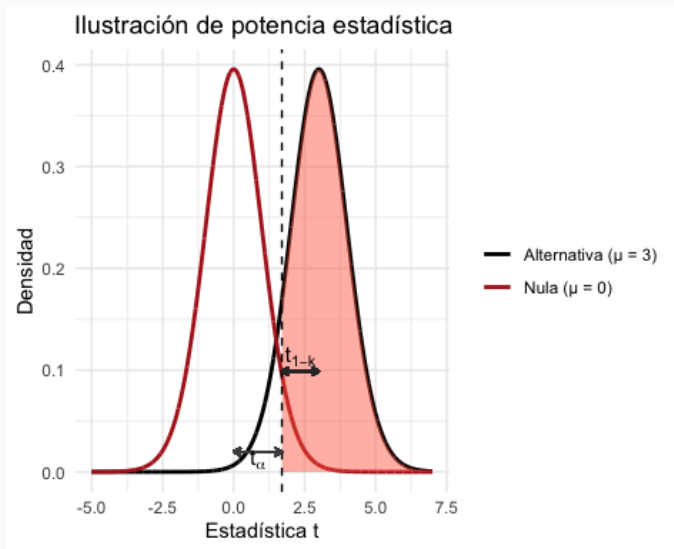
En ese caso, la varianza del estimador $\hat{\beta}$ es:

$$\frac{1}{P(1 - P)} \frac{\sigma^2}{N}$$

Vamos suponer que queremos una potencia de al menos κ . Normalmente se usa 80%.

Podemos preguntar: ¿qué efectos reales podemos identificar con al menos esse nivel de potencia?

Prueba de Hipótesis en Ensayos Aleatorios



Para obtener potencia κ , el parámetro tiene que ser:

$$\beta > (t_{1-\kappa} + t_{\alpha})se(\hat{\beta})$$

Para $\kappa = 80\%$, $t_{1-\kappa} = 0.84$.

Por lo tanto, tenemos una probabilidad de 80% de rechazar la nula si el efecto real es de aprox 2.8 veces el SE.

Así podemos definir el **Efecto Mínimo Detectable**.

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

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

Supón que tenemos $\alpha = 5\%$, $\kappa = 80\%$, $P = 50\%$, $N = 144$ y $\sigma = 1$. ¿Cual es el EMD?

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

- Existe un tradeoff entre el nivel de significancia y la potencia: si elegimos α más pequeño (t_{α} más grande), tenemos menos probabilidad de rechazar una nula verdadera, pero también no podemos detectar efectos pequeños.

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

Si deseamos maximizar la potencia (minimizar EMD), que debemos elegir como la proporción de tratados, P ?

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

Si deseamos maximizar la potencia (minimizar EMD), que debemos elegir como la proporción de tratados, P?

Si N es fijo, tenemos más potencia si $P = 50\%$.

Prueba de Hipótesis en Ensayos Aleatorios

Pero, supón que existen custos diferentes para un control y un tratamiento. Si queremos maximizar la potencia bajo un presupuesto, tenemos:

$$\min_{N,P} \frac{1}{NP(1-P)}$$

Bajo:

$$Nc_c + PNc_t = B$$

Resulta:

$$\frac{P}{1-P} = \sqrt{\frac{c_c}{c_t}}$$

Prueba de Hipótesis en Ensayos Aleatorios

En la practica, es muy comun utilizar formulas como esa, o similares, para calcular el tamaño minimo de la muestra para se detectar determinado efecto.

Por ejemplo, en investigación en educación, podemos querer investigar algun tipo de intervención con un ensayo aleatorio. Podemos calcular el efecto promedio obtenido en intervenciones similares y computar el N necesario para detectarlo.

En educación particularmente, es comun que los efectos estimados sean pequeños, e necesitan muestras muy grandes para investigarlos.

Prueba de Hipótesis en Ensayos Aleatorios

Investigaciones con baja potencia tienen un problema paradoxal.

Por definición, si el potencia es baja, existe una alta probabilidad de que se encuentre un efecto nulo, mismo si el efecto existe.

Pero, siempre existe un 5% de probabilidad de que vamos encontrar un efecto significativo, por suerte.

Si los errores son grandes, ese 5% se pasa cuando $\hat{\beta}$ es muy grande por variaciones aleatorias. Si se publica los falsos positivos, pero no los resultados nulos, podemos pensar que existe un efecto muy grande, por que nuestras estimativas son poco precisas.

Errores Agrupados

Errores Agrupados

Imaginate que te contratan para avaliar el efecto de una campaña publicitaria de una marca de cerveza.

La empresa hice un experimento aleatorio con 120 individuos, de que la mitad vio la campaña. Su interés es si compraran más de su producto.

Utilizas la formula que discutimos y obtienes que hubo un efecto positivo y es significativo con p-valor 3%.

Después te informan que la aleatorización fue solamente entre dos tiendas: todos los 60 controles estaban en la tienda de Concepción, y los 60 tratados estaban en la tienda de Valparaiso.

¿Es un problema para tu inferencia?

El problema aquí es que la aleatorización no fue entre individuos, pero entre **grupos**.

Y es natural suponer que existe correlación dentro de cada grupo.

Por ejemplo: Puede ser que estaba más caliente en Valparaíso en el día de la encuesta, o que llovió en Concepción. Esos factores aleatorios van a afectar todos los tratados o todos los controles, y tener un grande efecto.

Entonces, la estimativa del error estandar no está buena.

En muchos tipos de ensayos aleatorios, la aleatorización es en un nivel más largo que el individuo.

- La comuna, la escuela, etc

Todos dentro del mismo grupo tienen la misma asignación de tratamiento.

Normalmente, en ese tipo de contexto existen interacciones entre las unidades dentro de un grupo. Los errores no son independientes.

Tenemos que ajustar nuestros errores para reflejar eso.

Vamos escrever un modelo con j un grupo e i un individuo.

$$Y_i = \alpha + \beta D_i + v_i + w_{ij}$$

Donde v_i son errores comunes al grupo i , y w_{ij} son errores individuales. $Var(v_i) = \tau^2$,
 $Var(w_{ij}) = \sigma^2$

Existen J grupos, cada uno con n individuos.

En este caso, los errores estándar de $\hat{\beta}$ es:

$$se_g(\hat{\beta}) = \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{n\tau^2 + \sigma^2}{nJ}}$$

Si la aleatorización fuera individual, los errores serían:

$$se_i(\hat{\beta}) = \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\tau^2 + \sigma^2}{nJ}}$$

Podemos calcular la relación entre los errores estándar bajo aleatorización grupal y individual:

$$\frac{se_g(\hat{\beta})}{se_i(\hat{\beta})} = D = \sqrt{1 + (n - 1)\rho}$$

Donde ρ es la **correlación intragrupal**, esto es, la parte de la variación total explicada por la variación grupal.

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

En el caso de la campaña publicitaria, supón que el valor de correlación intragrupal es de solo 6%. ¿Cuánto más grandes son los errores estándar que computamos asumiendo asignación individual?

En el caso de la campaña publicitaria, supón que el valor de correlación intragrupal es de solo 6%. ¿Cuánto más grandes son los errores estándar que computamos asumiendo asignación individual?

$$D = \sqrt{1 + 59 * 0.06} \approx 2.1$$

¡Los errores son más que dos veces mayores que pensamos!

Cuando calculamos el Efecto Mínimo Detectable en un ensayo agrupado, obtenemos:

$$EMD = (t_{1-\kappa} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}} \sigma$$

Podemos ver que:

- El número de participantes por grupo, n , tiene un efecto relativamente débil.
- El número de grupos es más importante: aprendemos más de un individuo en un nuevo grupo que de más un individuo en un grupo ya observado.
- La correlación intragrupal aumenta el EMD, particularmente cuando n es grande.

Inferencia Exata

La prueba de hipótesis clásica es válida asintóticamente, o sea, si N es grande.

Pero a veces tenemos experimentos pequeños, entre 20 y 40 unidades. La inferencia puede no ser válida para ese caso.

Otro problema común es que existan valores extremos y muy influyentes. Eso también puede generar problemas para la inferencia clásica.

Cuando trabajamos con ensayos aleatorios, podemos lidiar con esos problemas de otra forma.

En inferencia clasica, asumimos que las unidades son observaciones de una población más grande. Existe variabilidad aleatoria en la **muestreaje** de la población.

En inferencia exacta, no vamos pensar en una población abstrata. Vamos lidiar solamente con las unidades que existen. La variabilidad aleatoria viene **del diseño del ensayo**. O sea, la única aleatoriedad es la asignación del tratamiento.

Inferencia Exata - Ejemplo

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

Inferencia Exata - Ejemplo

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

$$ATE = 34/4 - 30/4 = 8.5 - 7.5 = 1$$

Inferencia Exata - Ejemplo

Queremos saber si ese efecto de 1 es consistente con variación aleatoria.

Entonces, precisamos de una hipótesis sobre o que pasaria si la asignación fuera diferente.

No podemos observar Y_0 para los tratados, ni Y_1 para los controles. Pero, podemos utilizar una hipótesis para reconstruirlos.

La Hipótesis Nula Exacta de Fisher es de que el efecto de tratamiento es siempre cero.

$$H_0 : \delta_i = 0$$

Diferente de la hipótesis tradicional: *el efecto médio es cero.*

Inferencia Exata - Ejemplo

Bajo H_0 , podemos completar la tabla

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

Inferencia Exata - Ejemplo

Bajo H_0 , podemos completar la tabla

Name	D	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	1	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Ahora podemos analizar que pasaría con el efecto estimado si la asignación del tratamiento fuera diferente.

Podemos generar otra asignación posible y calcular el efecto resultante.

Inferencia Exata - Ejemplo

Name	\widetilde{D}_2	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Inferencia Exata - Ejemplo

Name	\widetilde{D}_2	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

$$\widetilde{ATE} = 36/4 - 28/4 = 9 - 7 = 2$$

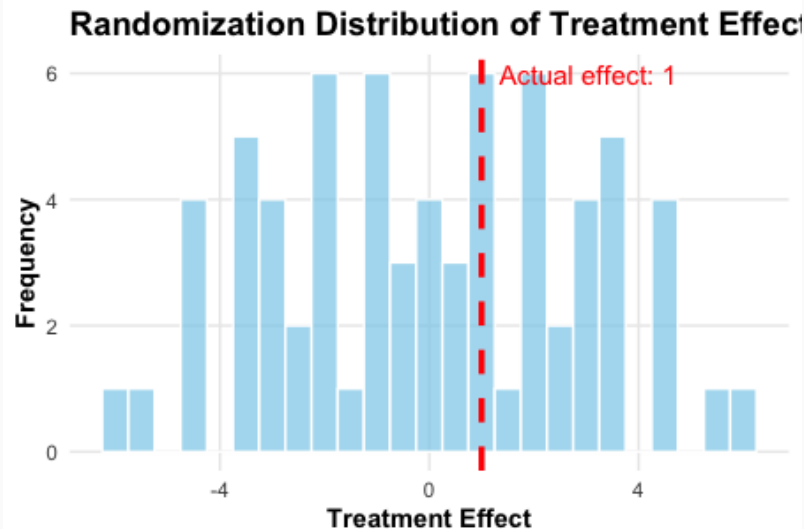
Inferencia Exata - Ejemplo

Podemos calcular el ATE bajo H_0 para **todas** las realizaciones posibles de la asignación del tratamiento.

Como tenemos 4 tratados de 8 individuos, las posibilidades son:

$$\frac{8!}{4!4!} = \frac{8 * 7 * 6 * 5}{4 * 3 * 2 * 1} = 70$$

Inferencia Exata - Ejemplo



Finalmente, podemos calcular el p-valor exato, por la proporción de realizaciones más extremas que el efecto real.

En la práctica, para cualquier aplicación un poco mayor, existe un número muy grande de posibilidades, y no podemos calcular todas.

Entonces, lo que hacemos es reasignar el tratamiento muchas veces de manera aleatoria, como una aproximación de la distribución completa.

Note que no usamos distribuciones estadísticas o argumentos basados en $n \rightarrow \infty$.
Cuando la muestra es pequeña, ese tipo de inferencia funciona mejor.

Si nos existen outliers con valores muy influentes, podemos utilizar otra estadística, como el rank médio:

- Hacemos el mismo proceso, pero en vez de computar el efecto promedio $(\frac{1}{n_t} \sum D_i Y_i - \frac{1}{n_c} \sum (1 - D_i) Y_i)$, computamos la diferencia en su rank $(\frac{1}{n_t} \sum D_i R_i - \frac{1}{n_c} \sum (1 - D_i) R_i)$.
- Donde R_i es una variable igual a cuantas unidades tienen Y_j menor que Y_i .

Otra ventaja es que podemos utilizar el conocimiento de la regla de asignación.

Por ejemplo, si tenemos asignación por grupos, podemos simplemente incorporar esa información cuando calculamos las otras asignaciones posibles.

Usando Covariadas como Controles

Podemos pensar en cuatro tipos de variables de controle.

Las necesarias

Son las que necesitamos incluir para cerrar todos los caminos backdoor. Si no las incluimos, nuestra investigación no es valida.

Ej: estratos si la asignación depende de estratos

Podemos pensar en cuatro tipos de variables de controle.

Las buenas

Son las que no son necesarias, pero ayudan a predecir Y , y por tanto disminuyen la varianza de los residuos.

No cambian la identificación del parámetro que estamos estimando, pero ayudan a obtener estimaciones más precisas. Aumentan la potencia de las pruebas.

Ej: el resultado de interés medido antes del tratamiento normalmente es un buen predictor.

Podemos pensar en cuatro tipos de variables de controle.

Las malas

No son necesarias, y predicen mejor el tratamiento que el resultado. Por eso, *disminuyen* la precisión.

No son comunes en ensayos aleatorios, pero a veces es posible incluir controles demasiados.

Podemos pensar en cuatro tipos de variables de controle.

Las prohibidas

Son las que cierran caminos causales importantes, o abren caminos de backdoor.

Ej: Colliders. Muchas veces resultados que ocurren después del tratamiento (controlar por ocupación después de una capacitación).