# Prueba de hipótesis

Luis Maldonado, PUC Chile

Agosto, 2023

- El problema fundamental de la inferencia causal es que los valores factuales y potenciales de la variable resultado no pueden ser observados al mismo tiempo para cada unidad.
- Por lo tanto, si para Luis se produce un efecto causal contrafactual del tratamiento $T$, cuando $y_{Luis,T=1} \neq y_{Luis,T=0}$, entonces ¿Cómo podemos aprender sobre el efecto causal?
- Una solución es la estimación de los promedios de los efectos causales (ATE).
- Esto es lo que llamamos el enfoque de Neyman.

- Otra posible solución es hacer afirmaciones o suposiciones sobre los efectos causales.
- Podríamos decir: "Creo que el efecto sobre Luis es 5" o "Este experimento no ha tenido ningún efecto sobre nadie". Y entonces podríamos preguntarnos "¿Cuánta evidencia tiene este experimento sobre esa afirmación?"
- Esta evidencia se resume en un valor $p$.
- A esto lo llamamos enfoque de Fisher.
- El enfoque de las pruebas de hipótesis para la inferencia causal nos dice cuánta evidencia o información obtenemos del diseño de la investigación sobre una afirmación causal.

## Componentes de una prueba de hipótesis

- Una **hipótesis** es una afirmación sobre una relación entre variables de resultado potenciales.
- Una **estadística de prueba** resume la relación entre el tratamiento y las variables de resultado observadas.
- El **diseño** nos permite vincular la hipótesis y la estadística de prueba: podemos calcular una estadística de prueba que describa una relación entre variables de resultado potenciales.
- El **diseño** también nos indica cómo generar una distribución de las posibles estadísticas de prueba sugeridas por la hipótesis.
- Un valor **p** describe la relación entre nuestra estadística de prueba observada y la distribución de las posibles estadísticas de prueba hipotéticas.

**Una hipótesis es una afirmación o modelo de una relación entre posibles variables de resultado**

| Outcome | Treatment | $y_{i,0}$ | ITE | $y_{i,1}$ | $Y > 0$ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 10 | 10 | 0 |
| 30 | 1 | 0 | 30 | 30 | 0 |
| 0 | 0 | 0 | 200 | 200 | 0 |
| 1 | 0 | 1 | 90 | 91 | 0 |
| 11 | 1 | 1 | 10 | 11 | 0 |
| 23 | 1 | 3 | 20 | 23 | 0 |
| 34 | 1 | 4 | 30 | 34 | 0 |
| 45 | 1 | 5 | 40 | 45 | 0 |
| 190 | 0 | 190 | 90 | 280 | 1 |
| 200 | 0 | 200 | 20 | 220 | 1 |

Por ejemplo, la hipótesis nula de ausencia de efectos, débil o estricta, es $H_0 : y_{i,1} = y_{i,0}$

## Las estadísticas de pruebas resumen las relaciones entre el tratamiento y las variables de resultado

```
## La estadística de prueba de diferencia de medias
meanTT <- function(ys, z) {
  mean(ys[z == 1]) - mean(ys[z == 0])
}

observedMeanTT <- meanTT(ys = Y, z = T)
observedMeanTT
[1] -49.6
```

## El diseño conecta la estadística de prueba y la hipótesis

- Lo que observamos para cada persona $i$ $(Y_i)$ es lo que habríamos observado en el tratamiento $(y_{i,1})$ o lo que habríamos observado en la situación de control $(y_{i,0})$.

$$Y_i = T_i y_{i,1} + (1 - T_i) * y_{i,0}$$

- Entonces, si $y_{i,1} = y_{i,0}$ por lo tanto $Y_i = y_{i,0}$.
- Lo que realmente observamos es lo que habríamos observado en la condición de control.

## El diseño guía la creación de una distribución de estadísticas de prueba hipotéticas

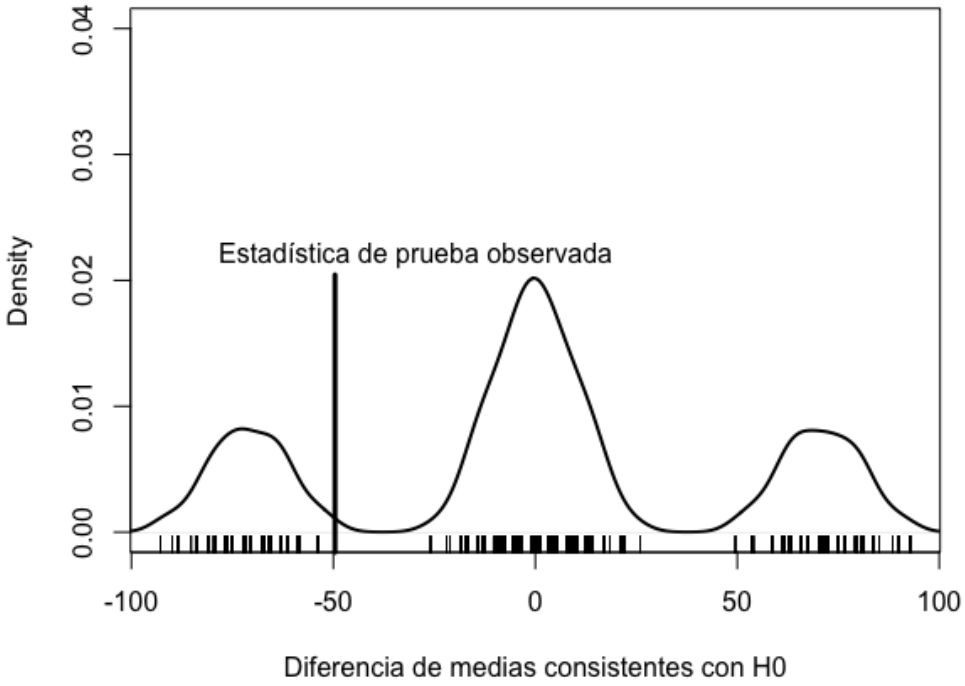
```
# Necesitamos saber cómo repetir nuestro experimento:

repeatExperiment <- function(N) {
  complete_ra(N)
}

# Luego lo repetimos, calculando la estadística de prueba implícita
# por la hipótesis y el diseño cada iteración:

set.seed(123456)
possibleMeanDiffsH0 <- replicate(
  10000,
  meanTT(ys = Y, z = repeatExperiment(N = 10))
)
```

# Planear las distribuciones de aleatoriedad bajo la hipótesis nula



## Los valores $p$ resumen los planes

¿Cómo debemos interpretar los valores $p$? (Nótese que son de una cola)

```
pMeanTT <- mean(possibleMeanDiffsH0 >= observedMeanTT)
pMeanTT
[1] 0.7785
```

- El valor $p$ es la probabilidad de obtener un test estadístico al menos tan grande como el test estadístico observado, dado que la hipótesis nula es verdadera.
- Nuestro ATE estimado es -49.6 y el valor $p$ es 0.7785, entonces la probabilidad de obtener una estimación igual o mayor a -49.6 simplemente por chance o suerte es de 78%.

## Cómo hacer esto en R: COIN

```
library(coin)
set.seed(12345)
pMean2 <- coin::pvalue(oneway_test(Y ~ factor(T),
                                   data = dat,
                                   distribution = approximate(nresample = 1000),
                                   alternative = "less"))

pMean2
[1] 0.783
99 percent confidence interval:
 0.7476049 0.8156543
```

## Cómo hacer esto en R: RI2

¿Cómo deberíamos interpretar el valor p de dos colas aquí?

```
## usando el paquete ri2
library(ri2)
thedesign <- declare_ra(N = N)

dat$Z <- dat$T

pMean4 <- conduct_ri(Y ~ Z,
  declaration = thedesign,
  sharp_hypothesis = 0, data = dat, sims = 1000
)

summary(pMean4)
 term estimate two_tailed_p_value
1    Z    -49.6          0.4444444
```

## Probando las hipótesis nulas débiles de que no hay efectos promedio

- La hipótesis nula débil es una afirmación sobre los agregados, y casi siempre se plantea en términos de promedios: $H_0 : \bar{y}_1 = \bar{y}_0$
- La estadística de prueba para esta hipótesis es casi siempre la diferencia simple de medias (por ejemplo, el "meanTT()" anteriormente mencionado).
- ¿Por qué el valor $p$ de OLS es diferente? ¿Qué supuestos utilizamos para calcularlo?

```
# OLS
lm1 <- lm(Y ~ T, data = dat)
summary(lm1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    78.20      33.97   2.302   0.0503 .
T             -49.60      48.04  -1.032   0.3321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
# t test, SE de Neyman
library(estimatr)
difference_in_means(Y ~ T, data = dat)
Design:  Standard
  Estimate Std. Error  t value  Pr(>|t|)  CI Lower CI Upper       DF
T    -49.6   48.04477 -1.03237 0.3587401 -181.5649 82.36491 4.112655
```

```
## SE de Neyman a mano (usado por difference_in_means)
# Funcion
varEstATE <- function(Y, T) {
  var(Y[T == 1]) / sum(T) + var(Y[T == 0]) / sum(1 - T)
}

# SE
seEstATE <- sqrt(varEstATE(dat$Y, dat$T))

# t de student
obsTStat <- observedMeanTT / seEstATE

# Valor p
c(
  observedTestStat = observedMeanTT,
  stderror = seEstATE,
  tstat = obsTStat,
  pval = 2 * min(
    pt(obsTStat, df = 8, lower.tail = TRUE),
    pt(obsTStat, df = 8, lower.tail = FALSE)
  )
)
```

```
observedTestStat          stderror             tstat              pval
     -49.6000000        48.0447708        -1.0323704         0.3320959
```

- El error estándar es la desviación estándar de una distribución muestral.

- El error estándar es útil como medida de la incerteza de la estimación del parámetro de interés (por ejemplo, ATE). Sabemos que la varianza de estimaciones OLS es

$$var(\hat{\beta}) = \frac{\sigma^2}{SST(1 - R^2)}. \tag{1}$$

- con $SST = \sum_{i=1}^{n}(x_i - \bar{x})$. Error estándar es la raiz cuadrada de (1).

- A medida que aumenta el error estándar, aumenta la incerteza en torno al parámetro estimado. Por lo tanto, una pregunta clave es ¿cómo reducir el error estándar?

- Considere un experimento donde tenemos $N$ unidades $i$. $M$ unidades son asignadas al grupo de tratamiento y $N-M$ son asignadas al grupo de control.
- Si $Z$ designa el tratamiento y $Y$ es la variable resultado, sabemos que los resultados observados están dados por

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0). \tag{2}$$

- El estimador de diferencias de medias (difference in means estimator) de ATE es

$$\beta = \bar{Y}(1) - \bar{Y}(0) = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)). \tag{3}$$

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} (Y_i) - \frac{1}{N-M} \sum_{i=M+1}^{N} (Y_i). \tag{4}$$

- El estimador de la varianza exacta de $\hat{\beta}$ es

$$var(\hat{\beta}) = \frac{N}{N-1}\left[\frac{\sigma^2_{Y(1)}}{M} + \frac{\sigma^2_{Y(0)}}{N-M}\right]$$
$$+\frac{1}{N-1}\left[2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)}\right]$$

- El estimador de la varianza exacta de $\hat{\beta}$ es

$$var(\hat{\beta}) = \frac{N}{N-1}\left[\frac{\sigma^2_{Y(1)}}{M} + \frac{\sigma^2_{Y(0)}}{N-M}\right] + \frac{1}{N-1}\left[2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)}\right] \quad (5)$$

- Gerber y Green (2012) escriben (5) de la siguiente forma

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m}\right] + 2cov(Y_i(0), Y_i(1))}$$

donde $m$ es el número de tratados.

$$SE(\hat{ATE}) = \sqrt{\frac{1}{N-1}\left[\frac{mVar(Y_i(0))}{N-m} + \frac{(N-m)Var(Y_i(1))}{m}\right] + 2cov(Y_i(0), Y_i(1))}$$

- En base a esta ecuación, para reducir el error estándar tenemos que tener en cuenta los siguientes elementos:

  1. Tamaño del grupo bajo estudio ($N$), tamaño del grupo de tratamiento ($m$) y tamaño del grupo de control ($N-m$): a mayor $N$, el error estándar es más bajo.
  2. Varianzas de $Y_i(1)$ y $Y_i(0)$: mientras más pequeñas las varianzas, menor es el error estándar.
  3. Covarianza entre $Y_i(1)$ y $Y_i(0)$: a menor covarianza, menor es el error estándar.

$$var(\hat{\beta}) = \frac{N}{N-1}\left[\frac{\sigma^2_{Y(1)}}{M} + \frac{\sigma^2_{Y(0)}}{N-M}\right]$$
$$+\frac{1}{N-1}\left[2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)}\right]$$

- Para estimar el error estándar del ATE estimado necesitamos los siguientes términos:
  - $\sigma^2_{Y(1)}$ o $Var(Y_i(1))$.
  - $\sigma^2_{Y(0)}$ o $Var(Y_i(0))$.
  - $cov(Y_i(0), Y_i(1))$.
- Las varianzas $Var(Y_i(1))$ y $Var(Y_i(0))$ son estimadas utilizando las varianzas de los grupos de tratamiento y control.

- Sabemos que los dos resultados potenciales para una misma unidad no pueden ser observados al mismo tiempo. En consecuencia, el término $cov(Y_i(0), Y_i(1))$ no puede ser estimado sin sesgo.

- Con un poco de matemática se puede demostrar que $2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)} \leq 0$. Si asumimos que $2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)} = 0$, tenemos un estimador del error estándar de ATE estimado conocido como *estimador de Neyman*:

$$\hat{SE} = \sqrt{\frac{Var(Y_i(0))}{N-m} + \frac{Var(Y_i(1))}{m}} \tag{6}$$

- Decimos que (6) es un estimador *conservador* del error estándar de ATE en el sentido de que es lo más grande que puede ser el verdadero error estándar, lo que puede traducirse en valores muy grandes.

- En el contexto de una regresión OLS, podemos estimar (6) con una versión del error estándar robusto a heteroscedasticidad llamada HC2.

- Comparemos el SE bajo homoscedasticidad y el SE robusto:

$$var(\hat{\beta}) = \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{7}$$

$$var(\hat{\beta}) = \frac{1}{n} \times \left[ \frac{\frac{1}{n-2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \right].$$

$$var(\hat{\beta}) = \frac{1}{n} \times \left[ \frac{\frac{1}{n-2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}{\left[ \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \right]. \tag{8}$$

- Existe una versión de (8), la cual se llama *HC2*.
- Ajuste de HC2: $\frac{\hat{u}_i^2}{1-h_{ii}}$ sustituye a $\hat{u}_i^2$ en (8). El término $h_{ii}$ es el *leverage* de la observación $i$ y tiene la siguiente expresión:

$$h_{ii} = \begin{cases} \frac{1}{m} & \text{if } T_i = 1 \\ \frac{1}{N-m} & \text{if } T_i = 0 \end{cases}$$

- con $m$ = tamaño del grupo de tratamiento. Un $h_{ii}$ grande significa que la observación $i$ tiene un fuerte impacto sobre el valor predicho de dicha unidad $i$.

- HC2 es equivalente al estimador de Neyman (Sammii and Aronow (2012)):

$$\hat{SE} = \sqrt{\frac{Var(Y_i(0))}{N-m} + \frac{Var(Y_i(1))}{m}}$$

- EL SE HC2 es consistente en muestras finitas e infinitas. Además, no asume ni homoscedasticidad ni linealidad.

- Otros estimadores robustos a la heteroscedasticidad (por ejemplo, (8)) son asintóticamente equivalentes a HC2.

- Recuerde corrección de Neyman: Tanto el SE de Neyman como HC2 son *conservadoramente sesgados*. El sesgo es

$$-\frac{1}{N-1}\left[2cov(Y(1), Y(0)) - \sigma^2_{Y(1)} - \sigma^2_{Y(0)}\right] \geq 0. \qquad (9)$$

- "Típicamente, el nivel de una prueba [$\alpha$] es una promesa sobre el rendimiento de esta, el tamaño es un dato sobre su rendimiento..." (Rosenbaum 2010, Glosario)
- $\alpha$ es la probabilidad de rechazar la hipótesis nula cuando la hipótesis nula es verdadera.
- ¿Qué significa "rechazar" $H_0 : y_{i,1} = y_{i,2}$ con un $\alpha = .05$?

## Errores de falsos positivos y falsos negativos

- **Error de falso positivo**: afirmamos que detectamos algo, pero en realidad no hay señal, solo ruido. Por ejemplo, decimos que hay un efecto ( rechazamos la hipótesis nula), pero en realidad éste es cero.
- **Error de falso negativo**: afirmamos que no podemos detectar algo, pero en realidad sí hay una señal. Por ejemplo, decimos que no hay un efecto (no rechazamos la hipótesis nula), pero en realidad sí lo hay.

- Un buen test o prueba debería:
  1. Raramente arrojar dudas sobre la verdad.
  2. Distinguir claramente la señal del ruido.

- Podemos utilizar nuestro diseño para saber si el procedimiento de prueba utilizado tiene bajo control los errores: hacer un diagnóstico.

- ¿Cómo se puede hacer el diagnóstico?: 1) Usar simulación, 2) Análisis de poder/potencia.

- Gerber, A. S. y D. P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton. Caps 3 y 4

- Discusión sobre test de hipótesis:
  - Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." Statistical Science 23 (2): 219–36.
  - Hodges, J. L., and E. L. Lehmann. 1963. "Estimates of location based on rank tests." Ann. Math. Statist 34: 598–611.
  - Rosenbaum, Paul R. 1993. "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies." Journal of the American Statistical Association 88 (424): 1250–53.
  - Rosenbaum, Paul R. 2010. "Design of observational studies." Springer Series in Statistics.

- Discusión sobre SE en experimentos:
  - Freedman, David (2008). On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics*, 40: 180-193.
  - Sammii, Cyrus y Peter M. Aronow (2012). On Equivalencies between Design-Based and Regression-Based Variance estimators for Randomized Experiments. *Statistics and Probability Letters* 82: 365-370.
  - Lin, Winston (2013). Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *The Annuals of Applied Statistics*, 7(1):295-318.