Taller de Test de hipótesis

José A. Ruiz-Tagle

20 de octubre, 2023

Antecedentes

En estudios experimentales, la randomización es esencial para garantizar la equivalencia inicial de los grupos y, por ende, proporcionar validez a las inferencias causales resultantes del experimento. Esta técnica implica la asignación aleatoria de sujetos a diferentes condiciones o tratamientos, minimizando así cualquier sesgo o factor confusor En este contexto, el paquete `blockrand` de R ofrece una herramienta específica para la randomización por bloques, permitiendo una asignación balanceada de tratamientos en grupos predeterminados.

El objetivo de este trabajo será utilizar test de hipótesis para determinar si existen diferencias significativas entre el grupo de tratamiento y control. Como investigadores, esperamos que los grupos no presenten diferencias significativas entre todos los factores observados.

Descripción de la base de datos La base `diabetes.csv` contiene variables basales sobre pacientes con y sin diabetes.

- Pregnancies: Número de embarazos
- Glucose: Nivel de Glucosa en sangre
- BloodPressure: Presión arterial
- SkinThickness: Indicador del grosor de la piel
- Insulin: Nivel de insulina en la sangre
- BMI: Índice de Masa Corporal(IMC)
- Age: Edad
- DiabetesPedigreeFunction: Porcentaje de diabetes
- Outcome: 1=Tiene diabetes, 2= No tiene

Randomización

```
data <- read.csv("https://github.com/JoseRTM/Clases-R/raw/main/diabetes.csv")

# Creamos la variable id
data$id <- 1:nrow(data)
n_participantes <- nrow(data)

# Definir el tamaño de bloque
tamano_bloque <- 4

# Definir los grupos de tratamiento
grupos <- c("Tratamiento", "Control")
```

```r
# Realizar la randomización por bloques
asignacion <- blockrand(n_participantes, num.levels = 2, levels = grupos)
# Ver la asignación

data <- data %>% dplyr::inner_join(asignacion, by = "id")

head(data)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome id block.id block.size   treatment
## 1                    0.627  50       1  1        1          4 Tratamiento
## 2                    0.351  31       0  2        1          4     Control
## 3                    0.672  32       1  3        1          4     Control
## 4                    0.167  21       0  4        1          4 Tratamiento
## 5                    2.288  33       1  5        2          4     Control
## 6                    0.201  30       0  6        2          4 Tratamiento
```

```r
# Reestructuramos los datos para las pruebas t
# Vamos a seleccionar solo las columnas relevantes antes de pivotar.
data_relevant <- data %>%
  select(-id, -block.id, -Outcome, -block.size) # Excluye las columnas que no quieres pivotar.

# Ahora, pivotamos el dataframe modificado
data_long <- data_relevant %>%
  pivot_longer(
    cols = -treatment, # Esto pivotará todas las columnas excepto 'treatment'.
    names_to = "variable",
    values_to = "value"
  )


# Realizamos pruebas t para cada variable y creamos una tabla resumen
resumen <- data_long %>%
  group_by(variable) %>%
  summarise(
    # Realizamos la prueba t y extraemos directamente el intervalo de confianza
    ci_data = list(broom::tidy(t.test(value ~ treatment, data = cur_data()))),
    media_Tratamiento = mean(value[treatment == "Tratamiento"], na.rm = TRUE),
    sd_Tratamiento = sd(value[treatment == "Tratamiento"], na.rm = TRUE),
    media_Control = mean(value[treatment == "Control"], na.rm = TRUE),
    sd_Control = sd(value[treatment == "Control"], na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  # Aquí, vamos a 'desempacar' el intervalo de confianza y prepararlo para la tabla
  mutate(
    ci_data = purrr::map(ci_data, ~ .x %>% select(conf.low, conf.high)),
    ci_lower = purrr::map_dbl(ci_data, ~ .x$conf.low),
    ci_upper = purrr::map_dbl(ci_data, ~ .x$conf.high),
```

variable	Grupo de tratamiento		Intervalo_Confianza
	Tratamiento	Control	
Age	33.15 (11.66)	33.33 (11.87)	[-1.49, 1.84]
BMI	31.98 (7.59)	32.01 (8.18)	[-1.09, 1.15]
BloodPressure	70.01 (18.81)	68.21 (19.87)	[-4.54, 0.94]
DiabetesPedigreeFunction	0.47 (0.33)	0.48 (0.33)	[-0.04, 0.05]
Glucose	119.18 (31.59)	122.61 (32.30)	[-1.10, 7.95]
Insulin	81.95 (121.99)	77.65 (108.20)	[-20.64, 12.03]
Pregnancies	3.84 (3.47)	3.85 (3.27)	[-0.48, 0.48]
SkinThickness	20.91 (16.27)	20.16 (15.64)	[-3.02, 1.51]

```
    Tratamiento = sprintf("%.2f (%.2f)", media_Tratamiento, sd_Tratamiento),
    Control = sprintf("%.2f (%.2f)", media_Control, sd_Control),
    Intervalo_Confianza = sprintf("[%.2f, %.2f]", ci_lower, ci_upper)
  ) %>%
  select(variable, Tratamiento, Control, Intervalo_Confianza) # Seleccionamos las columnas relevantes
resumen %>%
  kable("latex", booktabs = TRUE, align = "c") %>%
  kable_styling(latex_options = "striped", full_width = FALSE) %>%
  column_spec(1, bold = TRUE) %>%
  add_header_above(c(" " = 1, "Grupo de tratamiento" = 2, " " = 1))  # Ajusta según tus columnas
```

Objetivo: Interpretar los intervalos de confianza y contestar la siguiente pregunta:

- En base a toda la información disponible. ¿Usted diría que la randomización cumplió su objetivo? Es decir, son comparables los grupos? Se observan diferencias sustantivas entre el grupo de tratamiento y control?
- El primer párrafo debe interpretar todos los intervalos de confianza. La forma correcta de citar un intervalo de confianza en el texto es la siguiente: "Se observa una diferencia significativa entre X e Y (95%CI: X.XX;X.XX)."
- El segundo párrafo debe contener una conclusión que responde a las interrogantes planteadas.
- Escribir en máximo 1 plana en times new roman N°12 interlineado 1.5.
- Incluir los nombres de las personas que integran el grupo dentro y fuera del documento.

Ítem 2: Lectura

En base al texto "Contra la sumisión estadística" describa cuáles son las limitaciones de los test de hipótesis y de qué manera se pueden complementar. Debe fundamentar su respuesta en base al texto.