Análisis de Datos Categóricos

Ayudantía 4

Felipe Olivares

Contenido

Loops en R

Los loops ("ciclos" o "bucles" en español) son un tipo especial de funciones en R que sirven para ejecutar una tarea determinada una cantidad *n* de veces. Se llama iteración a cada una de estas repeticiones, y sirven para hacer en segundos lo que manualmente llevaría horas, días o sería simplemente demasiado.

```
# Sintaxis for

for (i in 1:4){
  print(i)
}
```

```
# [1] 1
# [1] 2
# [1] 3
# [1] 4
```

Loops en R

Componentes de un loop:

for = función que identifica los procedimientos del loop.

in = especificación del objeto (vector, factor, matriz) sobre el que se llevarán a cabo las iteraciones.

() = argumentos de unidades *i* o *j* de la función.

{} = operaciones de la función sobre cada *i* o *j*

```
#Sintaxis
for (i in 1:4){
print(i ^ 2)  # elevamos al cuadrado cada "i" que estamos iterando
}
```

```
# [1] 1
# [1] 4
# [1] 9
# [1] 16
```

Loops en R

Note que las especificaciones de un loop pueden ser aplicadas para la creación de distintos objetos, como es el caso de una matriz con filas *i* y columnas *j*. Por otro lado, se pueden definir vectores para integrarlos en los loops. En este sentido, losloop son muy flexibles respecto de lo que pueden realizar, ya sea con números o palabras

```r
# Sintaxis para vectores

perros <- c("Naruto", "Chopin", "Yeti", "sultan", "fido", "yonofui")

for (i in 1:length(perros)) {
  print(paste("Mi perro se llamaba:", perros[i]))
}
```

```
# [1] "Mi perro se llamaba: Naruto"
# [1] "Mi perro se llamaba: Chopin"
# [1] "Mi perro se llamaba: Yeti"
# [1] "Mi perro se llamaba: sultan"
# [1] "Mi perro se llamaba: fido"
# [1] "Mi perro se llamaba: yonofui"
```

Loops en R

Mediante loops es posible obtener resultados de promedios, medianas u otras operaciones que necesitemos de nuestros datos. Esto puede ser de mucha ayuda en la presencia bases de datos más grandes dónde, por ejemplo, sacar un promedio para cada valor es imposible.

```
# Usando la librería tidyverse generamos datos aleatorios en formato tibble
df <- tibble(
  a = rnorm(3),
  b = rnorm(3),
  c = rnorm(3),
  d = rnorm(3))

#resivamos los datos
df
```

```
## # A tibble: 3 x 4
##         a      b      c       d
##     <dbl>  <dbl>  <dbl>   <dbl>
## 1 -1.47   -1.67   -0.719  1.22
## 2  0.617   0.455   1.08  -0.994
## 3  0.725  -0.710  -0.346  0.0339
```

Loops en R

```r
#Sacar resultados 1 por 1
median(df$a)
```

```
# [1] 0.6171565
```

```r
median(df$b)
```

```
# [1] -0.7102614
```

```r
median(df$c)
```

```
# [1] -0.3463874
```

```r
median(df$d)
```

```
# [1] 0.03386827
```

```r
# Loops para los mismos resultados
resultados <- vector("double", ncol(df))  #  output
for (i in seq_along(df)) {                #  secuencia
  resultados[[i]] <- median(df[[i]])      #  cuerpo
}
resultados
```

```
# [1]  0.61715648 -0.71026144 -0.34638739  0.03386827
```

Loops en R

el comando *if* en loops permite colocar condiciones dentro de las operaciones que estamos realizando. Esto puede ser muy útil, por ejemplo, para definir dónde queremos realizar ciertas operaciones dentro de una base de datos. Veamos un ejemplo:

```
x <- 5
if(x > 0){ #condición
print("número positivo")
}
```

```
# [1] "número positivo"
```

Acá la expresión del código es "verdadera" si la sentencia se ejecuta, dada la condición que colocamos incialmente *"$x > 0$"*, por lo que generara una respuesta en "character" que dice "número positivo", lo cual definimos en el output.

Loops en R

Veamos un ejemplo más complejo. Supongamos que Juan tiene las siguientes notas en distintas asignaturas: 3,6,2,1,5,7 y queremos agregar una columna de aprobado y reprobado según la asignatura.

```
class <- c(3,6,2,1,5,7)  # notas de distintos ramos
p <- c()
for (i in 1:length(class)){
if(class[i] >= 4) p[i] = "aprobado"
if(class[i] < 4) p[i] = "reprobado"
}
as.data.frame(rbind(p,class))
```

```
#              V1       V2         V3         V4       V5        V6
# p     reprobado aprobado reprobado reprobado aprobado aprobado
# class         3        6          2          1        5         7
```

Donde los componentes *if*:

if = función de especificación lógica.

() = condiciones lógicas para unidades *i* o *j* de la función.

{} = operaciones de la función *if* sobre cada *i* o *j*

Medidas de asociación

Trabajaremos las medidas de asociación utilizando la base de datos del Observatoriod de conflictos (OCS) de COES. Esta base de datos es pública y la hemos estado utilizando en ayudantías anteriores. La base de datos está previamente trabajada para sus disintas variables, es decir, ha sufrido recodificaciones y creación de variables.

```
head(df1)
```

```
# # A tibble: 6 x 16
#   ano   region     educa~1 indig~2 laboral salud pacif~3 disru~4 viole~5 orga
#   <fct> <fct>      <fct>   <fct>   <fct>   <fct> <fct>   <fct>   <fct>   <fct
# 1 2009  Metropoli~ No      No      No      No    Sí      No      No      1 or
# 2 2009  Tarapacá   Sí      No      No      No    Sí      No      No      1 or
# 3 2009  Tarapacá   No      No      Sí      No    Sí      Sí      No      1 or
# 4 2009  O´Higgins  No      No      Sí      No    No      Sí      No      Sin
# 5 2009  Araucanía  No      Sí      No      No    Sí      No      No      Sin
# 6 2009  Araucanía  No      No      No      No    No      Sí      No      Sin
# # ... with 6 more variables: nacional <fct>, macrozona <chr>,
# #   estudiantes <fct>, trabajadores <fct>, ppolicial <fct>, apolicial <fct>,
# #   and abbreviated variable names 1: educacion, 2: indigena, 3: pacifica,
# #   4: disruptiva, 5: violenta, 6: organizacion
# # i Use ‘colnames()‘ to see all variable names
```

Medidas de asociación

Contexto:

En la última década han existido desarrollos teóricos importantes para aclarar el significado de la represión policial y para recentrar su estudio en torno a nociones más amplias de "control de la protesta", o "el control social de la protesta". En relación con esto último, Earl (2004) sostiene que parte relevante de la investigación sobre el control de la protesta estudia la represión como forma característica de observar las conductas que tienen las policías en escenarios de movilización social (Koopman, 1995; Kriesi et al, 1995; Mc Adams, 1982, Davenport. 2007). Este tipo aproximaciones basadas en las acciones represivas de las policías para mantener el orden público muchas veces oscurecen interrogantes importantes respecto de la heterogeneidad de los actores y conductas policiales empleadas. Por un lado, nublan la posibilidad de observar otras acciones que no involucran el uso dela coerción como estrategia de disuasión o control del orden público. Por otra parte, hablar de control dela protesta permite desmontar la idea de que la represión es un sinónimo de violencia estatal y esclarecerel arco de posibilidades que existen en las conductas policiales durante las manifestaciones.

En esta ocación estaremos trabajando sobre las medidas de asociación a través de la relación entre acción policial y tácticas o repertorios de la protesta. Esto datos podemos encontrarlos para distintas regiones y distintos años entre el 2009-2019. Para efectos del análisis utilizaremos la variable **acción** policial, la cuál fue creada a partir de distintas variables que contiene la base de datos del OCS (enfrentamientos directos con manifestantes, uso de carros lanzagua o gases lacrimógenos, uso de armas de fuego, uso de detenciónd de manifestantes, solo presencia policial de la protesta).

Medidas de asociación

```
# Presencia policial 2009-2019
table(df1$ppolicial) # 26% y 74%
```

```
#
#    Si    No
#  6079 17318
```

```
# Acción policial 2009-2019
table(df1$apolicial) # 44% y 56%
```

```
#
#  Control negociado Violencia Policial
#                2697                3382
```

```
# Tácticas de la protesta  2009-2019
table(df1$violenta) # 45% y 55%
```

```
#
#    No    Sí
# 19365  4033
```

Medidas de asociación

Lo primero que haremos es ver la asociación que existe entre las acciones policiales y las tácticas disruptivas durante una manifestación en el espacio público. Esto revisado de forma transversal para los datos que contiene el OCS para los años 2009-2019

```
# Sintaxis
#Relación bivariada entre presencia policial y táctica pacífica

ctable1 <- df1 %>% with(table(apolicial,disruptiva)) # 2-way table

print(ctable1)
```

```
#                     disruptiva
# apolicial                No   Sí
#   Control negociado    1248 1449
#   Violencia Policial   1365 2017
```

Medidas de asociación

Recordatorio:

Tenemos independencia estadística si la ocurrencia de un evento no afecta la probabilidad de la ocurrencia de otro evento. Dicho de otro modo, la probabilidad de que ocurra y es independiente de qué valor asume x. Por lo tanto, nos encontramos frente a independencia estadística si las probabilidades conjunta son iguales al producto de sus probabilidades marginales $P(XY) = P(X)P(Y)$. Las medidas de asociación nos permitirán justamente poder evaluar este escenario.

```
# Sintaxis
#Relaciones multivariadas (en este caso bivariadas)
prop.table(ctable1,1)
```

```
#                     disruptiva
# apolicial                   No        Sí
#   Control negociado  0.4627364 0.5372636
#   Violencia Policial 0.4036073 0.5963927
```

Medidas de asociación

¿Cuál es la diferencia de proporciones δ que se observa en la acción policial (Y) de acuerdo a la presencia de tácticas disruptivas de los manifestantes (X). Particularmente, la proporción de un control violento de la protesta (PCN) respecto de un control negociado de la protesta (PCV).

```
# Sintaxis
#Diferencia de proporciones
delta <- (0.597-0.537)
delta
```

```
# [1] 0.06
```

R= Existe una diferencia de proporción de 0.06 entre un control violento de la protesta para tácticas disruptivas en comparación con un control negociado de la protesta. Ahora, es importante recalcar que esto es solo una diferencia de proporciones, es decir, aún no sabemos si esto es estadísticamente significativo o no.

Medidas de asociación

Odds Ratio

La *odds* basícamente se define como la razón entre éxito o fracaso, es decir, la razón entre p y $1-p$ ¿Cuáles son las odds de un control violento de la protesta durante una manifestación disruptiva?

```
print(ctable1) # utilizamos la primera tabla de contigencia
```

```
#                      disruptiva
# apolicial                No   Sí
#   Control negociado    1248 1449
#   Violencia Policial   1365 2017
```

```
# Sintaxis
# Valor probabilístico de éxito
p <- (2017/(1365 +2017))

# Odds (P/1-P)
odds <- (p/(1-p))
odds
```

```
# [1] 1.477656
```

R:Las odds ("chances") de que existe un control violento de la protesta durante una manifestación disruptiva son de 1.47

Medidas de asociación

Odds ratio

A partir de la tabla de contigencia que tenemos podemos medir la asociación entre variables, es decir las *odds ratio*.

Ahora bien, en una tabla de 2x2 la razón de odds θ es la razón de éxito en dos filas, o $\theta = \frac{odds1}{odds2} = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}$

Sabemos que...

- Si $\theta = 1$ hay igualdad de odds ("chances") y, por lo tanto, hay independencia entre variables.
- Si $\theta > 1$ entonces el éxito es más probable para el grupo en el numerador.
- Si $\theta < 1$ entonces el éxito es más probable para el grupo en el denominador.

Medidas de asociación

Siempre es relevante, antes de calcular el *odds ratio*, saber que queremos calcular o que pregunta nos estamos haciendo. Por ejemplo:

¿Cuál es la razón de odds de un control negociado de la protesta en presencia de tácticas disruptivas?.

```
prop.table(ctable1,1)
```

```
#                     disruptiva
# apolicial                   No        Sí
#   Control negociado  0.4627364 0.5372636
#   Violencia Policial 0.4036073 0.5963927
```

```
# Sintaxis
# Razón de Odds en proporciones
OR <- ((0.537/0.462)/(0.596/0.403))
OR
```

```
# [1] 0.7859431
```

Medidas de asociación

```
print (ctable1)
```

```
#                     disruptiva
# apolicial                No   Sí
#   Control negociado  1248 1449
#   Violencia Policial 1365 2017
```

```
# esto también se puede calcular como producto cruzado

theta = ((1449*1365)/(1248*2017)) # Cross-product Ratio
theta
```

```
# [1] 0.7857431
```

```
theta_p <- (theta-1)*100
theta_p
```

```
# [1] -21.42569
```

R: Las odds de la existencia de un control negociado de la protesta son 0.8 veces las odds de un control violento de la protesta cuándo existen tácticas disruptivas, es decir, el control negociado de la protesta es un 21 % más bajo que el control violento de este tipo de manifestaciones con presencia de tácticas disruptivas.

Inferencia Estadística

Continuando con nuestros cálculos previos...

Intervalos de confianza

¿Es posible afirmar que nuestro valor $\delta = 0.06$, que observa en la muestra una diferencia de proporciones, es estadísticamente significativo a un 99% de confianza?

```
# Sintaxis
#CI Diferencia de proporciones
PCV  <- 0.5963927 # control violento
PCN  <- 0.5372636 # Control negociado
se <-  sqrt((PCV*(1 - PCV))/4690 +  (PCN*(1 - PCN))/3919)
ci99_delta <-  c(li=(delta - 2.58*se), ls=(delta + 2.58*se))
print(ci99_delta)
```

```
#         li         ls
# 0.03236132 0.08763868
```

R: La diferencia de proporciones observada es estadísticamente significativa a un 99% de confianza

Inferencia Estadística

Test χ^2

¿Es posible rechazar la H_0 que afirma la independencia estadística de la relación bivariada entre la presencia de tácticas disruptivas y el tipo de control policial de la protesta? *Ojo*: Recordemos que en tablas de 2x2 la independencia estadística entre variables equivale a $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, con una posible $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$ y que los grados de libertad están determinados por $df = (i-1)(j-1)$.

```
# Sintaxis
# Test Chi2 de independencia estadística
chisq.test(ctable1,correct = FALSE)
```

```
#
#   Pearson's Chi-squared test
#
# data:  ctable1
# X-squared = 21.405, df = 1, p-value = 3.718e-06
```

R= Con un valor $\chi^2 = 21.405$ y un $p = 0.000003718$ es posible rechazar H_0 y afirmar que no existe independencia estadística entre las variables en todos los niveles convencionales de confianza $p < 0.001$, $p < 0.01$, $p < 0.05$. Hay asociación entre las variables.