Análisis bivariado

Tema 1. Correlación simple: paramétrica y no paramétrica

Tema 2. Pruebas de hipótesis de correlación

Tema 3. Diferencias entre la correlación y la causalidad

Hasta este momento trabajamos describiendo una sola variable (análisis univariado) pero ¿cómo hacemos para determinar si existe alguna relación entre dos variables? Cuando buscamos identificar relaciones entre dos variables estamos frente a un análisis bivariado o bidimensional.

El tipo de resumen estadístico descriptivo siempre va a depender de la clasificación de las variables involucradas.

Para trabajar en esta unidad continuaremos utilizando el dataframe ***mydata.csv*** en el cual construiremos la variable GRUPO categorizando la variable EDAD cada 3 años.

```
> datos <- read.table( "m“data.csv","header = TRUE,
                sep = "," " dec = "."")"> datos$GRUPO <- cut(datos$EDAD, breaks = seq(20, 35,
by = 3),
                include.lowest = T)
```

01. Cualitativas vs cualitativas

Con la intención de reunir en una sola estructura toda la información disponible de dos variables categóricas creamos tablas de doble entrada, llamadas **tablas de contingencia**.

Construiremos una tabla de contingencia entre la variable SEXO y la variable GRUPO que creamos anteriormente. Para esto es necesario que ambas variables se encuentren en formato *factor*. En caso que esto no ocurra así, se debe forzar la clase a factor con la función ***factor()*** que vimos en el bloque anterior.

Para crear las tablas de contingencia usamos la función **table()** solo que debemos incorporar dos variables dentro de sus argumentos:

```
> table(datos$GRUPO, datos$SEXO)
          femenino masculino
 [20,23]        3         10
 (23,26]        8          4
 (26,29]        4          3
 (29,32]        5          2
 (32,35]        5          6
```

Los resultados devueltos son valores absolutos. Para agregar las sumatorias marginales, de fila y columna, utilizamos la función ***addmargins()***

```
> addmargins(table(datos$GRUPO,datos$SEXO))
femenino masculino Sum
 [20,23]        3         10  13
 (23,26]        8          4  12
 (26,29]        4          3   7
 (29,32]        5          2   7
 (32,35]        5          6  11
 Sum           25         25  50
```

Como paso siguiente haremos la tabla con valores porcentuales, para ello utilizaremos la función ***prop.table()*** como realizamos con las tablas univariadas.

Si queremos que el denominador sea el total fila, entonces deberemos incorporar el parámetro ***margin = 1.*** A su vez, todo estará dentro de la función ***round()*** para que los resultados se redondeen en 2 dígitos.

```
> round( 100 * (prop.table( table( datos$GRUPO, datos$SEXO),
            margin = 1)), 2)
femenino masculino
 [20,23]   23.08      76.92
 (23,26]   66.67      33.33
 (26,29]   57.14      42.86
 (29,32]   71.43      28.57
 (32,35]   45.45      54.55
```

En cambio, si queremos que el denominador sea por columna, igualamos ***margin=2***

```
> round( 100 * (prop.table( table( datos$GRUPO, datos$SEXO),
          margin = 2)), 2)
femenino masculino
  [20,23]      12       40
  (23,26]      32       16
  (26,29]      16       12
  (29,32]      20        8
  (32,35]      20       24
```

02. Cuantitativas vs cualitativas

Para relacionar variables cuantitativas con variable cualitativa, podemos aplicar los diferentes resúmenes de medidas para cada una de las categorías de la variable cualitativa.

Por ejemplo, podemos desear conocer cuál es la media de PESO para cada SEXO. En este caso resulta conveniente utilizar la función ***tapply().*** Esta función pertenece a la familia de apply y trabaja "aplicando" la función definida como argumento (en este caso mean) a un conjunto de datos según los factores de otra variable (femenino, masculino).

```
> tapply(datos$PESO, datos$SEXO, mean)

 femenino masculino
  55.6080   81.0356
```

Las funciones ***apply(), lapply(), vapply(), sapply()*** aplican, con algunas diferencias menores, funciones directamente sobre vectores, columnas de matrices y variables de dataframes, resultando en valores de tipo vector o lista.

La función ***tapply()*** no solo hace los cálculos por columna del dataframe, sino que también por las categorías que defina el usuario, dentro del grupo de observaciones.

Otras funciones que nos facilitan hacer la misma operación son ***by() y aggregate().***

by() trabaja de forma similar al uso de **apply()** pero la salida tiene un formato diferente:

```
> by(datos$PESO, datos$SEXO, mean)
```

```
>datos$SEXO: femenino
[1] 55.608
>datos$SEXO: masculino
[1] 81.0356
```

aggregate() se diferencia del resto porque tiene una estructura distinta: diremos que queremos agregar una variable en función de la otra. Para ello, es necesario introducir por un lado la relación (en este caso PESO en función del SEXO), luego el nombre del dataframe en el parámetro ***data*** y por último, la función que aplicaremos para agregar los datos en el parámetro ***FUN***.

```
> aggregate(PESO ~ SEXO, data = datos, FUN = mean)
    SEXO      PESO
1 femenino 55.6080
2 masculino 81.0356
```

Lo interesante de esta función es que podemos incorporar otras variables categóricas:

```
> aggregate(PESO ~ GRUPO + SEXO, data = datos, mean)
    GRUPO     SEXO    PESO
1  [20,23]  femenino 62.44333
2  (23,26]  femenino 55.50000
3  (26,29]  femenino 49.32500
4  (29,32]  femenino 55.03600
5  (32,35]  femenino 57.27800
6  [20,23] masculino 84.03700
7  (23,26] masculino 80.05000
8  (26,29] masculino 82.22333
9  (29,32] masculino 78.86000
10 (32,35] masculino 76.82167
```

03. Cuantitativas vs cuantitativas

Cuando se realiza el estudio conjunto de dos variables cuantitativas, normalmente el objetivo es determinar si existe algún tipo de asociación entre ellas o si, por el contrario, son independientes entre sí.

En términos prácticos, la asociación significa que el conocimiento de los valores de una de las variables proporciona alguna información sobre los valores de la otra.

03.01. Pruebas de hipótesis para correlación

Muchas veces necesitamos comprobar si una variable se correlaciona con otra, es decir si cuando una aumenta de valor la otra aumenta o disminuye de manera sistemática.

El análisis de correlación permite hallar un estadístico que es el **coeficiente de correlación de Pearson (r)**, el cual puede tomar valor que se hallan en el intervalo [-1,+1]. Cuanto más cercano sea a +1 indica una fuerte relación positiva entre las variables estudiadas, es decir cuando una aumenta, la otra también lo hace. Por su parte, un valor de r cercano a -1 muestra una gran correlación entre las variables, pero cuando una crece, la otra decrece. Mientras que, un valor cercano a 0, nos indica falta de relación entre las variables.

El coeficiente de correlación (r) elevado al cuadrado (r^2) se conoce como **coeficiente de determinación** y nos indica qué porcentaje de una variable es explicado por la variación de la otra.

En una correlación, también tendremos un p-valor.

Un valor de **p-valor < 0.05** nos indica que las variables **están correlacionadas**, mientras que un **p-valor > 0.05** indicará **falta de correlación**.

03.02. Tipos de correlación

- **Pearson**: evalúa el grado de relación entre dos variables cuantitativas.
- **Spearman**: evalúa el grado de relación entre dos variables de rango ordenado
- **Kendall:** es una medida no paramétrica de un rango de correlación.

03.03. Prueba de correlación de Pearson (r)

Permite determinar si existe asociación lineal entre dos variables continuas. Sus supuestos de aplicación son:

La muestra es un subconjunto aleatorio de la población.

No existe una asociación no lineal entre las variables.

Las variables presentan una distribución normal.

La prueba de correlación de Pearson permite aplicar el siguiente test de hipótesis:

Ho: r = 0. No existe asociación lineal entre las variables.

Ha: r ≠ 0. Hay asociación lineal entre las variables.

Continuamos con ejemplo anterior. Se decide trabajar con un nivel de confianza del 95% y un nivel de significación de α=0.05.

Para probar la significancia estadística hacemos uso de:

Hipótesis nula	Ho: No hay relación.
Hipótesis alternativa	Ha: Hay relación.

La función **cor.test** (x, y, alternative = , method =) nos ayuda a encontrar el nivel de significancia.

Donde

x = variable continua explicativa

y = variable continua respuesta

alternative = referencia a las colas: less, greater, two.side

method = tipo de correlación: pearson, spearman, kendall

Si queremos saber si la hipótesis de prueba es: la correlación de la población es menor a cero, usamos "less". Si queremos saber si es mayor a cero usamos: "greater". Y si solo queremos saber si es diferente de cero, usamos "two.side"

Para aplicar las distintas funciones, trabajaremos con las variables PESO y EDAD de nuestro dataframe ***"mydata.csv".***

El paquete base de R trae incorporada la función ***cov()*** para el cálculo de la **covarianza**:

```
> cov(datos$PESO,datos$EDAD)
[1] -17.75536
```

y la función ***cor()*** para el **coeficiente de correlación**:

```
> cor(datos$PESO,datos$EDAD)
[1] -0.241473
```

La función ***cor()*** utiliza por defecto el coeficiente de correlación de Pearson. Si queremos utilizar otro parámetro de correlación, debemos cambiar el parámetro **method**, el cual acepta también **"kendall"** o **"spearman"** como atributos.

Para saber si esta correlación es significativa debemos evaluar el p-valor. Para ello es conveniente utilizar la función ***cor.test()***

El coeficiente de correlación, r, es un estimador del parámetro poblacional ρ. Podemos preguntarnos si ciertamente existe una correlación entre los valores X y los valores Y de la población, y probar Ho: ρ = 0.

Mediante la función **cor.test()** de R, podemos realizar el análisis de correlación simple paramétrico (Pearson) y no-paramétricos (Kendall, Spearman). La correlación paramétrica provee la prueba de hipótesis Ho:ρ = 0 mediante el estadístico t, el intervalo de confianza (95 %) para r, y el valor de r.

Utilizando cor.test también podemos realizar las pruebas no-paramétricas, o correlaciones de rangos de Kendall y Spearman. La interpretación se hace examinando el valor de p, la probabilidad de cometer error Tipo I

```
> cor.test(datos$PESO,datos$EDAD)
        Pearson's product-moment correlation
data:  datos$PESO and datos$EDAD
t = -1.724, df = 48, p-value = 0.09115
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.48708221  0.03953182
sample estimates:
     cor
-0.241473
```

Analicemos otro ejemplo. Carguemos la Tabla3, que contiene datos concentración de hematocritos, viscosidad de la sangre y cantidad de proteínas.

```
>Tabla3<- read_excel("Tabla3.xlsx")
>summary(Tabla3)
      hto             visc      proteínas
 Min.   :29.00   Min.   :3.000   Min.   :5.000
 1st Qu.:34.75   1st Qu.:3.175   1st Qu.:6.900
 Median :38.00   Median :3.400   Median :7.150
 Mean   :38.38   Mean   :3.475   Mean   :7.125
 3rd Qu.:41.25   3rd Qu.:3.725   3rd Qu.:7.825
 Max.   :49.00   Max.   :4.200   Max.   :8.100
```

Evaluemos la normalidad para cada variable

```
> shapiro.test(Tabla3$hto)

        Shapiro-Wilk normality test

data:  Tabla3$hto

W = 0.97339, p-value = 0.8901
 > shapiro.test(Tabla3$visc)
        Shapiro-Wilk normality test
data:  Tabla3$visc
W = 0.92413, p-value = 0.1965

> shapiro.test(Tabla3$proteínas)
        Shapiro-Wilk normality test
data:  Tabla3$proteínas
W = 0.90451, p-value = 0.09495
```

Observando los p-valor aceptamos que las tres variables de la tabla tienen distribución normal.

Plantemos el análisis de correlación para las variables hematocrito (hto) y viscosidad (visc)

```
>cor.test(Tabla3$visc,Tabla3$hto,alternative="greater",method="pearson",conf.level=0.95)
        Pearson's product-moment correlation
 data:  Tabla3$visc and Tabla3$hto
t = 18.892, df = 14, p-value = 1.163e-11
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.9532137 1.0000000
sample estimates:
      cor
0.9809458
```

El resultado no indica un alto significado estadístico de la correlación (p-valor < 0,05) y alta correlación entre las variables, r = 0.98

Probemos con la variable proteína y viscosidad

```
>cor.test(Tabla3$visc,Tabla3$proteínas,alternative="two.sided",
        method="pearson", conf.level=0.95)

        Pearson's product-moment correlation
 data:  Tabla3$visc and Tabla3$proteínas
t = -1.7824, df = 14, p-value = 0.09638
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.76308494  0.08343862
sample estimates:
      cor
-0.4300542
```

Vemos que la correlación entre las variables es negativa, r= - 0.43, sin embargo ésta correlación no es significativa ya que p-valor > 0.05.

A continuación, analicemos si hay una relación lineal entre dos variables, pero sin asumir que una es funcionalmente dependiente de la otra, es decir, realicemos un análisis de correlación, usando el coeficiente de correlación (r), como el estadístico para medir dicha relación.

Imaginemos que se necesita evaluar la asociación entre el puntaje PSU (Prueba de Selección Universitaria -Chile) y la cantidad de libros por persona en el hogar, y que para ello, se cuenta

con una muestra aleatoria de estudiantes del país (n=2000). Decidimos trabajar con un nivel de confianza del 95% y un nivel de significación de α=0.05.

Uso la función **set.seed()** para generar datos aleatorios simulados (n=2000).

```
>set.seed(123) #fijar semilla aleatoria
>librospersona <- rnorm(2000, 5,2)
>librospersona[librospersona < 0] <- 0


>psu <- 200 + 20*librospersona + rnorm(2000,150,45)
```

Analizamos la dispersión de los datos a través de un gráfico.

```
>plot(librospersona, psu)
>abline(lm(psu ~ librospersona), col = "red")
```

```
> shapiro.test(librospersona)
        Shapiro-Wilk normality test
data:  librospersona
W = 0.99845, p-value = 0.06044


> shapiro.test(psu)
        Shapiro-Wilk normality test
data:  psu
W = 0.99948, p-value = 0.8901
```

Luego, evaluamos si se cumple el supuesto de normalidad en la distribución de las variables.

Dado que ambos p-valor > 0,05 se concluye que ambas variables siguen una distribución normal, comprobando los supuestos necesarios para realizar una prueba de correlación Pearson.

```
> cor.test(librospersona, psu,  method = "pearson", use = "complete.obs")
```

```
        Pearson's product-moment correlation
data:  librospersona and psu
t = 39.688, df = 1998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6387010 0.6877611
sample estimates:
      cor
0.6639449
```

Se observa un coeficiente de correlación Pearson (r) de 0.66, siendo estadísticamente significativo a un nivel de confianza de 95% (p-valor < 0,05). De esta manera, existe evidencia suficiente para afirmar que las personas que poseen más libros en su hogar son también quienes obtienen mejores puntajes PSU

03.04. Prueba de correlación de Spearman (ρ)

Permite determinar si existe asociación lineal entre dos variables continuas, cuando no se cumple el supuesto de normalidad y/o estamos trabajando con variables ordinales.

Sus supuestos de aplicación son:

La muestra es un subconjunto aleatorio de la población.

No existe una asociación no lineal entre las variables.

La prueba de correlación de Spearman permite evaluar la siguientes hipótesis nula y alternativa:

Ho: $\rho=0$. No existe asociación lineal entre las variables.

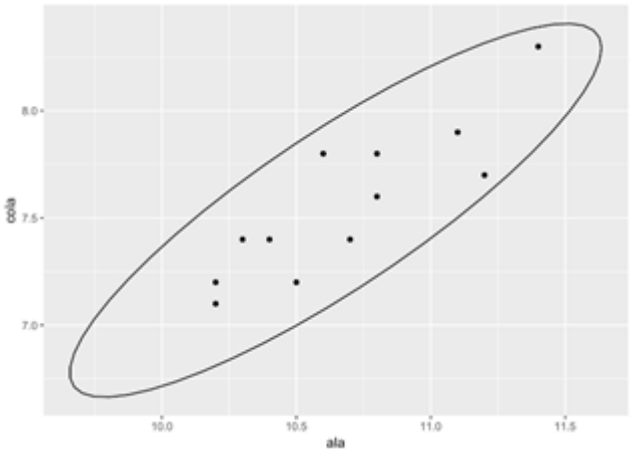
Ha: $\rho\neq0$. Hay asociación lineal entre las variables.

Otro ejemplo, vamos a calcular el coeficiente de correlación de Pearson, para la posible relación lineal entre la longitud del ala (cm) de un especie de ave, y el largo de su cola (cm) a partir de los siguientes datos.

```
>ala <- (10.4,10.8,11.1,10.2,10.3,10.2,10.7,10.5,10.8,11.2,10.6,11.4)
```

```
>cola <- c(7.4,7.6,7.9,7.2,7.4,7.1,7.4,7.2,7.8,7.7,7.8,8.3)

>scatter <- data.frame(ala,cola)
ggplot(scatter, aes(ala, cola)) +
  geom_point() +
  stat_ellipse()
```



Comencemos analizando la normalidad.

```
> shapiro.test(ala)
        Shapiro-Wilk normality test
data:  ala
W = 0.94168, p-value = 0.5201

> shapiro.test(cola)
        Shapiro-Wilk normality test
data:  cola
W = 0.94449, p-value = 0.5583
```

A continuación analicemos la correlación paramétrica.

```
> corranalp <- cor.test(ala,cola, method = "pearson")
> corranalp
        Pearson's product-moment correlation
```

```
data:  ala and cola
t = 5.5893, df = 10, p-value = 0.0002311
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5923111 0.9631599
sample estimates:
      cor
0.8703546
```

Ahora la correlación no paramétrica

```
> corranalk <- cor.test(ala,cola, method = "kendall")
> corranalk
        Kendall's rank correlation tau

data:  ala and cola
z = 3.1418, p-value = 0.001679
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.7202074
```

¿Cuál es tu conclusión?

03.05. ¿Por qué la correlación no implica causalidad?

La presencia de una correlación no siempre significa que haya una relación causal, así como la presencia de una relación causal no siempre significa que existe una correlación.

La correlación examina la relación entre dos variables. Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso solemos decir que "la correlación no implica causalidad".

```
> corranals <- cor.test(ala,cola, method = "spearman")
> corranals
        Spearman's rank correlation rho
data:  ala and cola
S = 42.59, p-value = 0.0004467
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8510852
```

La correlación examina la relación entre dos variables. Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso solemos decir que "la correlación no implica causalidad".