Análisis de correlación

J. Eduardo Gamboa U.

2025-10-13

Contents

1 Introducción **2**

2 Lectura y preprocesamiento de datos **2**

3 Asocación de variables cualitativas **3**
3.1 Ejemplo 1 3
3.2 Ejemplo 2 5

4 Asocación de variables cualitativas ordinales **7**
4.1 Ejemplo 7

5 Asociación de variable cualitativa con cuantitativa **9**
5.1 Ejemplo 1 9
5.2 Ejemplo 2 11

6 Asociación de variable cualitativa jerárquica con cuantitativa **12**
6.1 Ejemplo 1 12
6.2 Ejemplo 2 13

7 Asociación de variables cuantitativas **15**
7.1 Ejemplo 15

1 Introducción

Cuando trabajamos de forma bivariada (o multivariada), será de utilidad analizar la relación entre las variables. Las medidas y procedimientos para cuantificar esta asociación dependerán del tipo de variable que tengamos en análisis.

Para esta clase, utilizaremos el archivo `Employee.csv`, el cual fue publicado en Kaggle bajo el nombre IBM HR Analytics Employee Attrition & Performance.

Si bien el archivo contiene gran cantidad de variables, solo utilizaremos las siguientes:

- Attrition: Variable categórica binaria que indica si el empleado dejó la empresa (Yes / No)
- Department: Variable categórica referida al área de trabajo (Sales, Research & Development, etc.)
- BusinessTravel: Variable categórica ordinal sobre la frecuencia de viajes de negocios (Rarely, Frequently, etc.)
- Education: Variable categórica jerárquica que indica el nivel educativo: Below College, College, Bachelor, Master, Doctor.
- Age: Edad del empleado, en años.
- DistanceFromHome: Distancia que recorre el empleado desde casa.

2 Lectura y preprocesamiento de datos

```
library(dplyr)
library(ggplot2)

datos <- read.csv('Employee.csv')

set.seed(7)
datos |>
  select(Attrition, Department, BusinessTravel, Education,Age, DistanceFromHome) |>
  rename(Desercion  = 1, Depto = 2, Viaje = 3, Educacion = 4, Edad = 5, Distancia = 6) |>
  mutate(Viaje = factor(Viaje,
                        levels = c("Non-Travel", "Travel_Rarely", "Travel_Frequently"),
                        ordered = TRUE)) |>
  mutate(Educacion = case_when(Educacion == 1 ~ "Below College",
                               Educacion == 2 ~ "College",
                               Educacion == 3 ~ "Bachelor",
                               Educacion == 4 ~ "Master",
                               Educacion == 5 ~ "Doctor")) |>
  mutate(Educacion = factor(Educacion,
                            levels = c("Below College", "College", "Bachelor",
                                       "Master", "Doctor"),
                            ordered = TRUE)) |>
  slice_sample(n = 500) -> datos
```

```
datos |> head(5)
```

```
##   Desercion                  Depto             Viaje Educacion Edad Distancia
## 1        No Research & Development     Travel_Rarely    Master   47         9
## 2       Yes                  Sales Travel_Frequently  Bachelor   23         9
## 3        No                  Sales     Travel_Rarely   College   26        28
## 4        No                  Sales     Travel_Rarely    Doctor   39         2
## 5       Yes Research & Development     Travel_Rarely  Bachelor   29         1
```

3 Asocación de variables cualitativas

El interés consiste en determinar y cuantificar la relación entre dos variables categóricas, sean estas de tipo nominal o jerárquico. Para ello se emplea principalmente:

1. Tablas de contingencia

También llamadas tablas de doble entrada, son la herramienta fundamental para resumir y visualizar la distribución conjunta de las frecuencias de las categorías de dos o más variables cualitativas.

2. Gráfico de barras apiladas

Es un gráfico de barras tradicional donde cada barra se divide en segmentos que se colocan uno encima del otro (apilados).

3. Coeficiente V de Cramer

Es una medida general aplicable a tablas de contingencia. Sus valores varían entre 0 (independencia) y 1 (asociación perfecta). Cuando la tabla es 2×2, se tiene el coeficiente Φ

4. Prueba Chi Cuadrado de independencia

Es la prueba estadística más común. Contrasta la hipótesis nula H_0 de que las variables son independientes frente a la hipótesis alternativa (H_1) de que están asociadas.

3.1 Ejemplo 1

1. Tabla de contingencia:

```
tab1 <- table(datos$Desercion, datos$Depto)
tab1
```

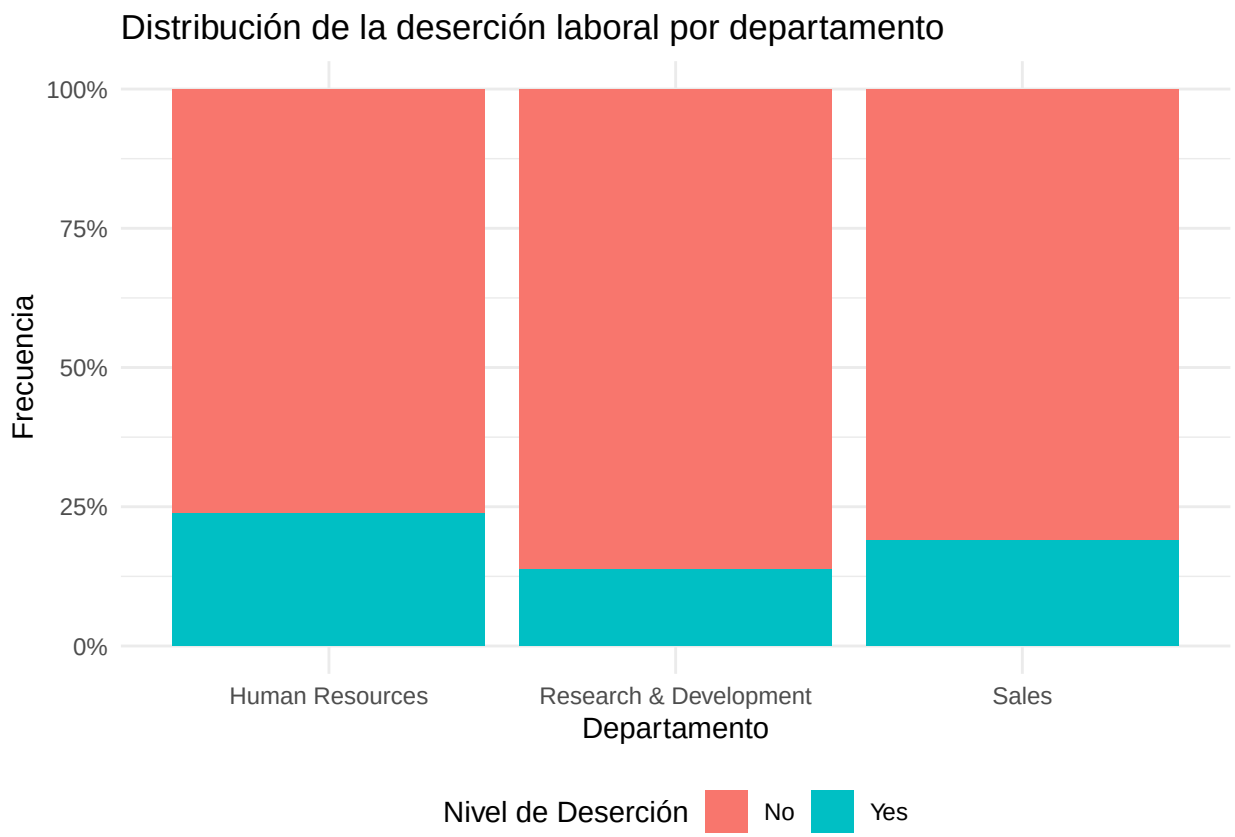
```
##
##       Human Resources Research & Development Sales
##   No               16                    276   129
##   Yes               5                     44    30
```

2. Gráfico de barras apiladas:

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.4.3
```

```
datos |>
  ggplot(aes(x = Depto, fill = Desercion)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Distribución de la deserción laboral por departamento",
       x = "Departamento",
       y = "Frecuencia",
       fill = "Nivel de Deserción") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



3: Coeficiente V de Cramer:

```
tab1 |> DescTools::CramerV()
```

```
## [1] 0.07935266
```

Su valor cercano a cero es una evidencia de que no existe asociación entre el departamento y la deserción laboral.

4. Prueba Chi Cuadrado de independencia:

H_0: El departamento y la deserción laboral son independientes

H_1: El departamento y la deserción laboral no son independientes

```
tab1 |> chisq.test()
```

```
## Warning in chisq.test(tab1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 3.1484, df = 2, p-value = 0.2072
```

Con un nivel de significancia del 5%, no se rechaza H_0, por lo tanto el departamento y la deserción laboral son independientes (no están asociados).

3.2 Ejemplo 2

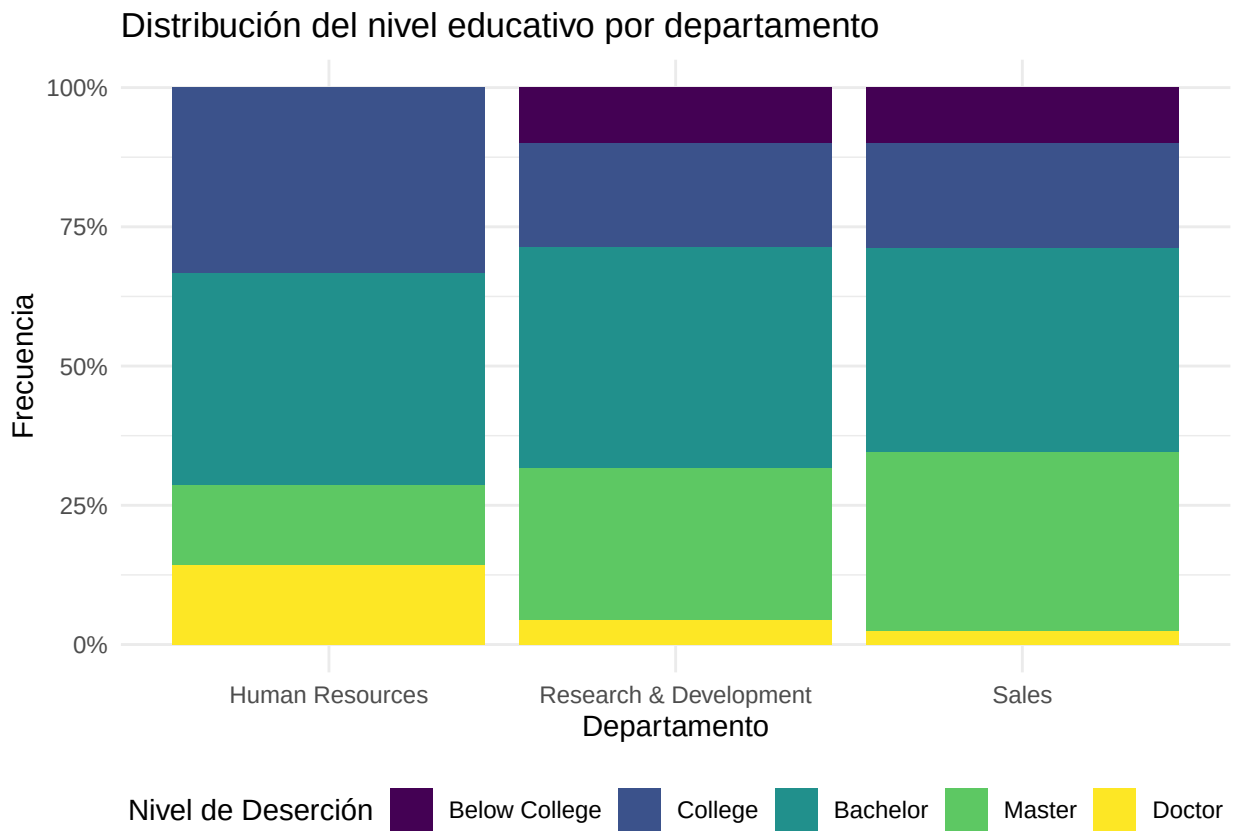
1. Tabla de contingencia:

```
tab2 <- table(datos$Educacion, datos$Depto)
tab2
```

```
##
##                  Human Resources Research & Development Sales
##   Below College                0                      32    16
##   College                      7                      60    30
##   Bachelor                     8                     127    58
##   Master                       3                      87    51
##   Doctor                       3                      14     4
```

2. Gráfico de barras apiladas:

```
datos |>
  ggplot(aes(x = Depto, fill = Educacion)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Distribución del nivel educativo por departamento",
       x = "Departamento",
       y = "Frecuencia",
       fill = "Nivel de Deserción") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



3: Coeficiente V de Cramer:

```
tab2 |> DescTools::CramerV()
```

```
## [1] 0.114771
```

Asociación muy baja entre el departamento y la deserción laboral.

4. Prueba Chi Cuadrado de independencia:

H_0: El departamento y el nivel educativo son independientes

H_1: El departamento y el nivel educativo no son independientes

```
tab2 |> chisq.test()
```

```
## Warning in chisq.test(tab2): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 13.172, df = 8, p-value = 0.1061
```

Con un nivel de significancia del 5%, no se rechaza H_0, por lo tanto el departamento y el nivel educativo son independientes (no están asociadas).

4 Asocación de variables cualitativas ordinales

1. Mapa de calor

Representa l distribución proporcional conjunta entre dos variables categóricas. Permite observar la distribución relativa de frecuencias dentro de cada nivel así como patrones de asociación o tendencias entre las variables.

2. Kendall τ-b:

Mide la fuerza y dirección de la asociación monótona entre dos variables ordinales. Se recomienda cuando las variables poseen un número similar de categorías (tabla de contingencia cuadrada o casi cuadrada) o cuando existen empates, situación frecuente en datos ordinales. Este coeficiente evalúa el grado de concordancia y discordancia entre los pares de observaciones, incorporando una corrección por empates para proporcionar una estimación más precisa de la relación. Fluctúa entre -1 y 1.

3. kendall τ-c (Stuart):

Mide la fuerza y dirección de la asociación monótona entre dos variables ordinales cuando éstas tienen diferente número de categorías (tabla de contingencia rectangular). A diferencia de τ-b, no corrige los empates, pero ajusta el valor máximo posible del coeficiente según el tamaño de la tabla, lo que permite comparar asociaciones entre tablas de distintas dimensiones. Fluctúa entre -1 y 1.

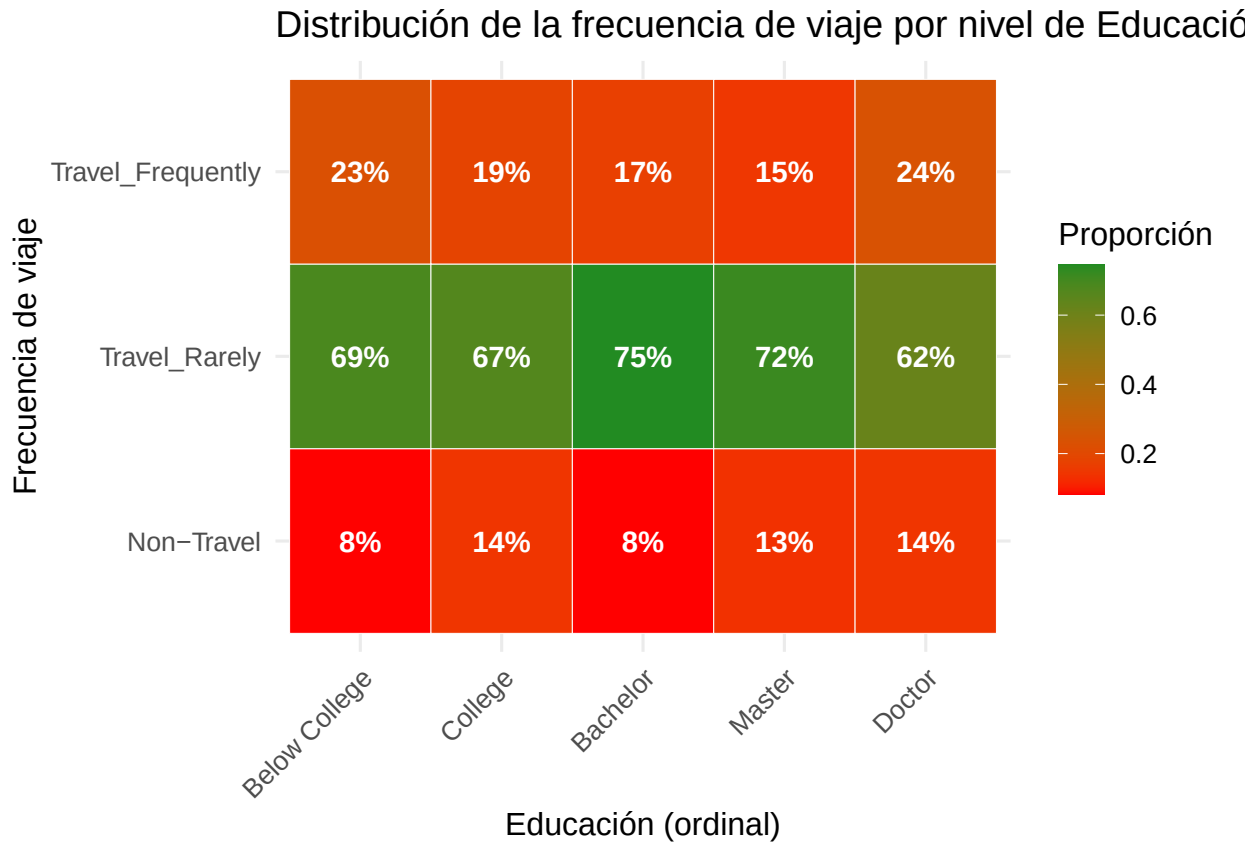
4. Coeficiente de correlación de Spearman:

Mide la fuerza y dirección de la relación monótona entre dos variables ordinales con muchos niveles o continuas. Maneja los empates promediando los rangos, pero si son muchos llega a haber un sesgo. Fluctúa entre -1 y 1.

4.1 Ejemplo

1. Mapa de calor

```
datos |>
  count(Educacion, Viaje) |>
  group_by(Educacion) |>
  mutate(prop = n / sum(n)) |>
  ungroup() |>
  ggplot(aes(x = Educacion, y = Viaje, fill = prop)) +
  geom_tile(color = "white") +
  geom_text(aes(label = percent(prop, accuracy = 1)), color = "white", fontface = "bold") +
  scale_fill_gradient(low = "red", high = "forestgreen", name = "Proporción") +
  labs(title = "Distribución de la frecuencia de viaje por nivel de Educación",
       x = "Educación (ordinal)", y = "Frecuencia de viaje") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



2. Kendall tau-b

```
datos$Viaje |> DescTools::KendallTauB(datos$Educacion, conf.level = 0.95)
```

```
##       tau_b      lwr.ci      upr.ci
## -0.03773725 -0.11934369  0.04386919
```

Valor cercano a cero → falta de asociación

3. Kendall tau-c (Stuart)

```
datos$Viaje |> DescTools::StuartTauC(datos$Educacion, conf.level = 0.95)
```

```
##        tauc      lwr.ci      upr.ci
## -0.03226800 -0.10211132  0.03757532
```

Valor cercano a cero → falta de asociación

4. Spearman

```
Edu <- datos$Educacion |> factor(ordered = TRUE) |> as.numeric()
Via <- datos$Viaje |> factor(ordered = TRUE) |> as.numeric()
Edu |> cor(Via, method = "spearman")
```

```
## [1] -0.04226009
```

Su valor cercano a cero indica falta de asociación, la cual además se confirma con el p-valor = 0.3457, mediante el cual no se rechaza $H_0 : \rho = 0$.

5 Asociación de variable cualitativa con cuantitativa

1. Boxplot

Representación gráfica que permite comparar la distribución de una variable cuantitativa entre los niveles de una variable cualitativa.

2. Coeficiente de correlación biserial puntual

Mide la fuerza y dirección de la relación entre una variable dicotómica real (0/1, Sí/No) y una variable cuantitativa continua. Toma valores entre -1 y 1, de modo que si el valor es cercano a 1, entonces el grupo "1 o Sí" tiene promedios más altos. De lo contrario, si el valor es cercano a -1, el grupo "1 o Sí" tiene promedios más bajos.

3. Análisis de varianza

Evalúa si existen diferencias estadísticamente significativas entre las medias de una variable cuantitativa en tres o más grupos definidos por una variable cualitativa.

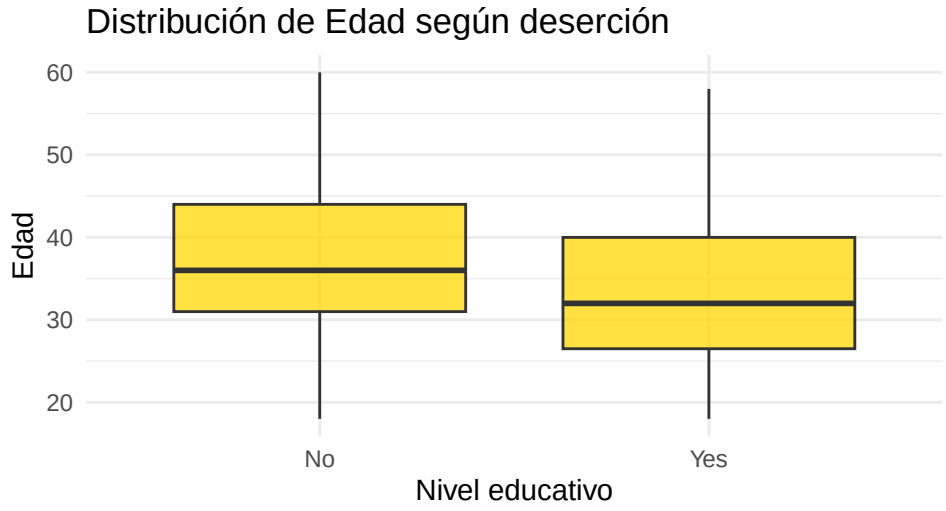
4. eta^2

Es un estadístico de tamaño de efecto derivado del ANVA. Es un estadístico de tamaño de efecto derivado del ANOVA. Si su valor es cercano a 0.01, la magnitud es pequeña. Alrededor de 0.06, mediana. Cercano a 0.14 es grande.

5.1 Ejemplo 1

1. Boxplot

```
datos |>
  ggplot(aes(x = factor(Desercion), y = Edad)) +
  geom_boxplot(fill = "gold", alpha = 0.75) +
  labs(title = "Distribución de Edad según deserción",
       x = "Nivel educativo", y = "Edad") +
  theme_minimal() +
  theme(legend.position = "none")
```



2. Coeficiente de correlación biserial puntual

```
ltm::biserial.cor(x = datos$Edad, y = datos$Desercion, level = 2)
```

```
## [1] -0.1466335
```

La asociación es baja, de modo que el grupo "Sí" tiene promedios ligeramente más bajos.

3. Análisis de varianza

```
m2 <- aov(Edad ~ Desercion, data = datos)
m2 |> summary()
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Desercion     1    956   955.9   10.94 0.00101 **
## Residuals   498  43501    87.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al menos una de las medias de Edad difiere entre los niveles de deserción. Como solo se trata de dos niveles, existe diferencia entre las medias de edad de los que desertaron y los que no.

4. η^2

```
m2 |> DescTools::EtaSq()
```

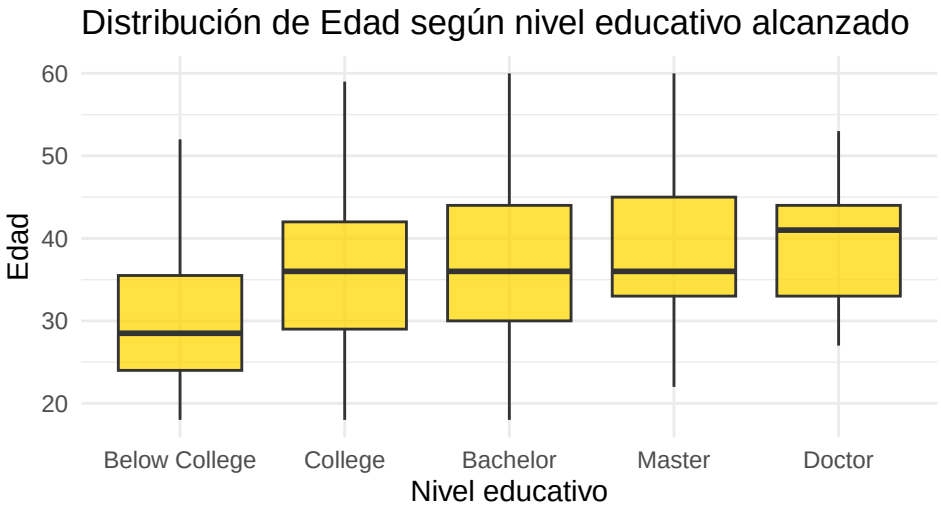
```
##               eta.sq eta.sq.part
## Desercion 0.02150137  0.02150137
```

La magnitud de asociación es baja.

5.2 Ejemplo 2

1. Boxplot

```
datos |>
  ggplot(aes(x = factor(Educacion, ordered = TRUE),
             y = Edad)) +
  geom_boxplot(fill = "gold", alpha = 0.75) +
  labs(title = "Distribución de Edad según nivel educativo alcanzado",
       x = "Nivel educativo", y = "Edad") +
  theme_minimal() +
  theme(legend.position = "none")
```



2. Análisis de varianza

```
m2 <- aov(Edad ~ Educacion, data = datos)
m2 |> summary()
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Educacion     4   3050   762.5   9.115 4.12e-07 ***
## Residuals   495  41407    83.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al menos una de las medias de Edad difiere entre los niveles de educación.

3. η^2

```
m2 |> DescTools::EtaSq()
```

```
##               eta.sq eta.sq.part
## Educacion 0.06860557  0.06860557
```

La asociación es de magnitud mediana.

6 Asociación de variable cualitativa jerárquica con cuantitativa

1. Boxplot

Representación gráfica que permite comparar la distribución de una variable cuantitativa entre los niveles de una variable cualitativa.

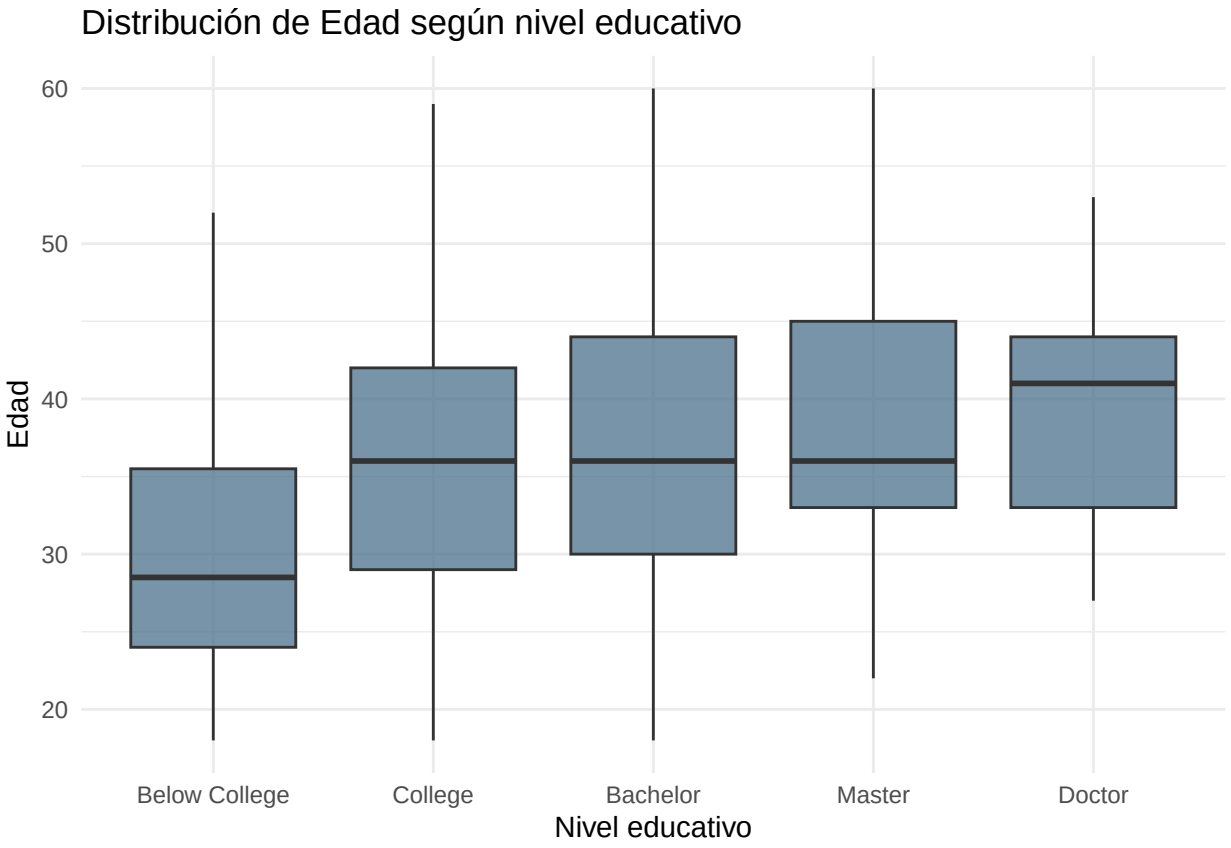
2. Coeficiente de correlación de Spearman:

Mide la fuerza y dirección de la relación monótona entre dos variables ordinales con muchos niveles o continuas. Maneja los empates promediando los rangos, pero si son muchos llega a haber un sesgo. Fluctúa entre -1 y 1.

6.1 Ejemplo 1

1. Boxplot

```
datos |>
  ggplot(aes(x = factor(Educacion, ordered = TRUE),
             y = Edad)) +
  geom_boxplot(fill = "skyblue4", alpha = 0.75) +
  labs(title = "Distribución de Edad según nivel educativo",
       x = "Nivel educativo", y = "Edad") +
  theme_minimal() +
  theme(legend.position = "none")
```



2. Coeficiente de correlación de Spearman:

```
cor(as.numeric(datos$Educacion), datos$Edad, method = "spearman")
```

```
## [1] 0.2267993
```

```
cor.test(as.numeric(datos$Educacion), datos$Edad, method = "spearman")
```

```
## Warning in cor.test.default(as.numeric(datos$Educacion), datos$Edad, method =
## "spearman"): Cannot compute exact p-value with ties
```

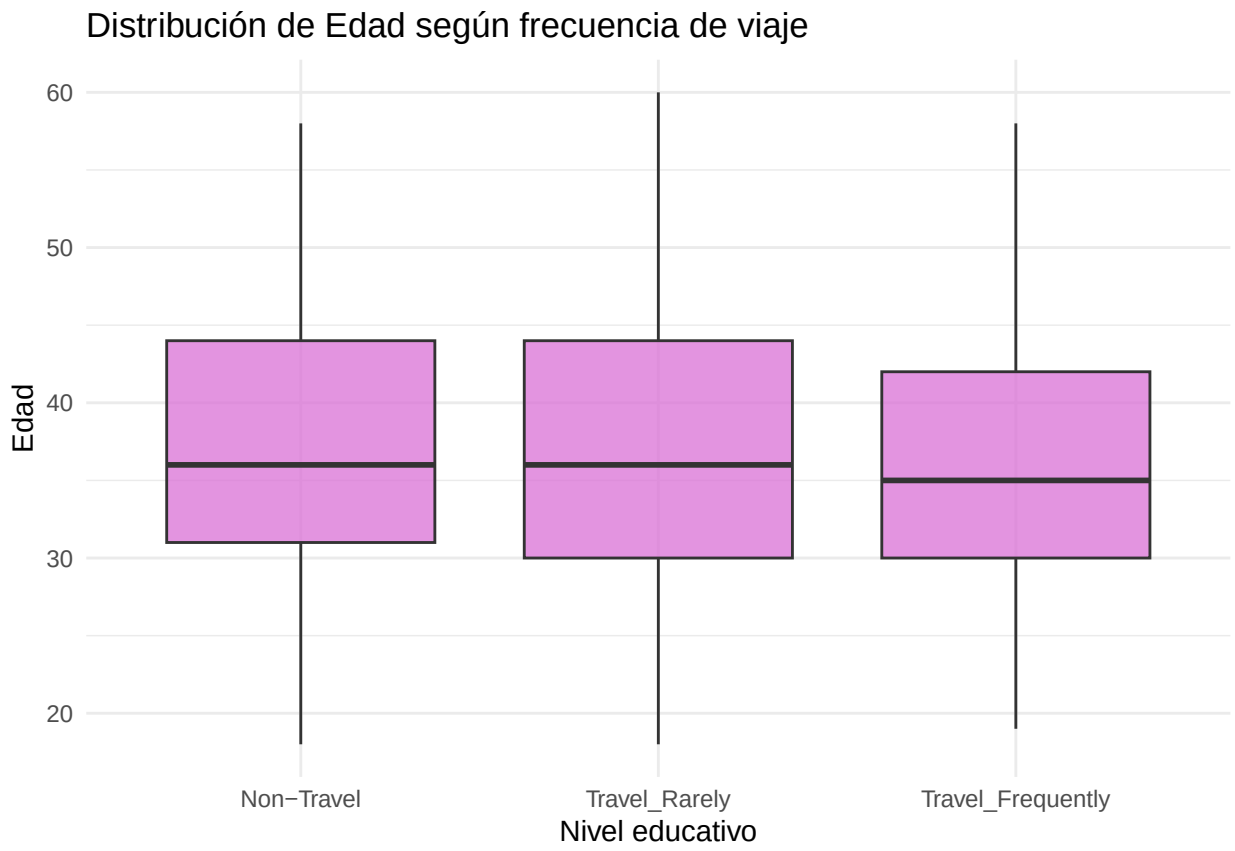
```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(datos$Educacion) and datos$Edad
## S = 16108284, p-value = 2.965e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2267993
```

La correlación es significativa, ya que el pvalor es cercano a cero. Además, la asocicación es positiva o directa pero débil, es decir a mayor edad, tiende a observarse mayor nivel educativo, aunque esta relación no es tan fuerte.

6.2 Ejemplo 2

1. Boxplot

```
datos |>
  ggplot(aes(x = factor(Viaje, ordered = TRUE),
             y = Edad)) +
  geom_boxplot(fill = "orchid", alpha = 0.75) +
  labs(title = "Distribución de Edad según frecuencia de viaje",
       x = "Nivel educativo", y = "Edad") +
  theme_minimal() +
  theme(legend.position = "none")
```



2. Coeficiente de correlación de Spearman:

```
cor(as.numeric(datos$Viaje), datos$Edad, method = "spearman")
```

```
## [1] -0.04111342
```

```
cor.test(as.numeric(datos$Viaje), datos$Edad, method = "spearman")
```

```
## Warning in cor.test.default(as.numeric(datos$Viaje), datos$Edad, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(datos$Viaje) and datos$Edad
## S = 21689776, p-value = 0.3589
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.04111342
```

La correlación no es significativa, ya que el pvalor es superior a 0.05. Así, no existe asociación entre la edad y la frecuencia de viaje.

7 Asociación de variables cuantitativas

1. Diagrama de dispersión

Representación gráfica que muestra la relación entre dos variables cuantitativas. Cada punto del gráfico representa una observación, ubicada según los valores de ambas variables. Permite identificar patrones de asociación, tendencias lineales o no lineales y valores atípicos.

2. Coeficiente de correlación de Pearson

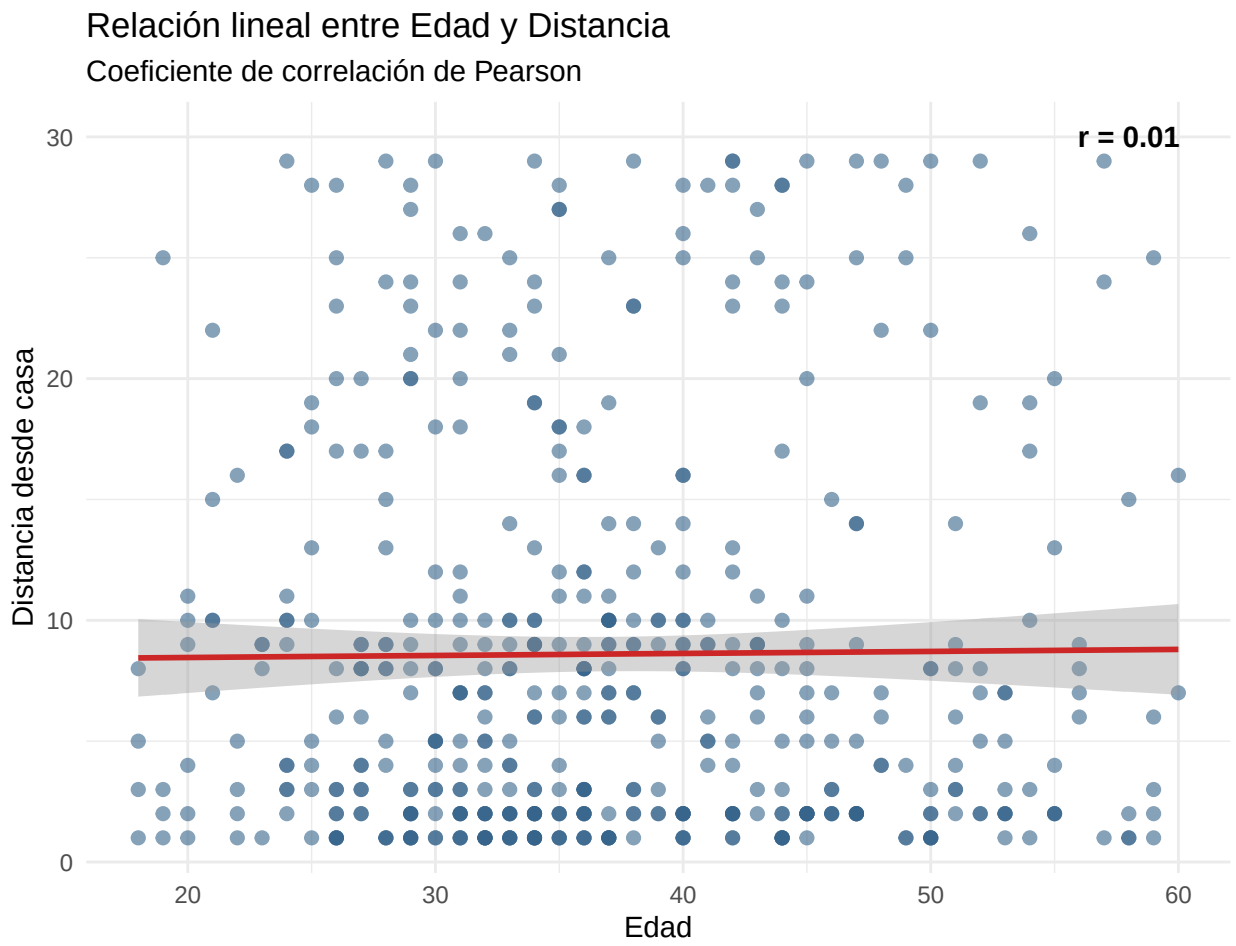
Mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas continuas. Su valor varía entre -1 y +1: valores positivos indican una relación directa, negativos una relación inversa, y cercanos a cero una relación lineal débil o inexistente.

7.1 Ejemplo

1. Diagrama de dispersión

```
r_pearson <- cor(datos$Edad, datos$Distancia, method = "pearson")
datos |>
  ggplot(aes(x = Edad, y = Distancia)) +
  geom_point(alpha = 0.6,
             color = "steelblue4",
             size  = 2) +
  geom_smooth(method = "lm",
              se     = TRUE,
              color  = "firebrick3") +
  annotate("text",
           x        = 58,
           y        = 30,
           label    = paste0("r = ", round(r_pearson, 2)),
           size     = 4.0,
           fontface = "bold") +
  labs(title    = "Relación lineal entre Edad y Distancia",
       subtitle = "Coeficiente de correlación de Pearson",
       x        = "Edad",
       y        = "Distancia desde casa") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



2. Coeficiente de correlación de Pearson

```
cor.test(datos$Edad, datos$Distancia, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$Edad and datos$Distancia
## t = 0.21789, df = 498, p-value = 0.8276
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07799389  0.09737070
## sample estimates:
##         cor
## 0.009763476
```

No existe asociación entre la edad y la distancia a casa, ya que el pvalor es alto (cercano a 1).