

# Factores

## HR Analytics: Teoría y Práctica

<http://pablohaya.com/contact>

09/2018

# Variables categóricas

Una variable es categórica cuando puede tomar valores de un conjunto finito.

Por ejemplo, el género (Masculino/Femenino), el grupo sanguíneo (O, A, B, AB), el estado civil (casado, soltero, divorcido, separado), las Comunidades Autónomas (Galicia, Andalucía, Cantabria...), o distintos tipos de departamento (RRHH, Marketing, Comercial, Administración...).

Las variables categóricas se definen mediante la función `factor()`

```
f <- factor(c("M", "F", "F", "M", "F"))
```

```
f
```

```
## [1] M F F M F
```

```
## Levels: F M
```

La función `summary()` nos permite obtener un recuento del número de valores que se ha encontrado por cada nivel:

```
f <- factor(c("M", "F", "F", "M", "F"))  
summary(f)
```

```
## F M  
## 3 2
```

# Variables ordinales

Una variable ordinal es una variable categórica en la que se puede establecer un orden entre los valores. Dentro de esta categoría entran un grado de satisfacción, una escala de puntuación Likert (totalmente desacuerdo... totalmente de acuerdo), el nivel de educativo, los meses del año...

Las variables categóricas se definen mediante el parámetro `ordered`

```
f <- factor(c("TD", "TA", "TA", "NN", "TA"), ordered = TRUE)
f
```

```
## [1] TD TA TA NN TA
## Levels: NN < TA < TD
```

Por defecto, el orden de las variables se asigna alfabéticamente. Se puede establecer el orden mediante añadiendo un vector que incluya los posibles valores (niveles) en el orden que queramos

```
f <- factor(c("TD", "TA", "TA", "NN", "TA"), ordered = TRUE,  
            c("TD", "NN", "TA"))
```

```
f
```

```
## [1] TD TA TA NN TA  
## Levels: TD < NN < TA
```

```
min(f)
```

```
## [1] TD  
## Levels: TD < NN < TA
```

Incluso es posible especificar niveles que no estuvieran en el vector original, pero que pudieran ser valores factibles.

```
f <- factor(c("TD", "TA", "TA", "NN", "TA"), ordered = TRUE,  
            c("TD", "D", "NN", "A", "TA"))
```

```
f
```

```
## [1] TD TA TA NN TA
```

```
## Levels: TD < D < NN < A < TA
```

```
summary(f)
```

```
## TD  D NN  A TA
```

```
##  1  0  1  0  3
```

Se puede renombrar los niveles originales dándoles otros nombres mediante el parámetro labels

```
f <- factor(c("TD", "TA", "TA", "NN", "TA"), ordered = TRUE,  
            c("TD", "D", "NN", "A", "TA"),  
            labels = c("Totalmente en desacuerdo",  
                        "En desacuerdo",  
                        "Ni de acuerdo ni en desacuerdo",  
                        "De acuerdo",  
                        "Totalmente de acuerdo"))
```

f

```
## [1] Totalmente en desacuerdo      Totalmente de acuerdo  
## [3] Totalmente de acuerdo           Ni de acuerdo ni en desa  
## [5] Totalmente de acuerdo  
## 5 Levels: Totalmente en desacuerdo < ... < Totalmente de ac
```

# levels

Si pueden cambiar los niveles sobre un factor ya creado mediante empleando la función `levels()`

```
f <- factor(c("TD", "TA", "TA", "NN", "TA"), ordered = TRUE,  
            c("TD", "D", "NN", "A", "TA"))  
levels(f)
```

```
## [1] "TD" "D"  "NN" "A"  "TA"
```

```
levels(f) <- c("Totalmente en desacuerdo", "En desacuerdo",  
               "Ni de acuerdo ni en desacuerdo", "De acuerdo",  
               "Totalmente de acuerdo")  
levels(f)
```

```
## [1] "Totalmente en desacuerdo"      "En desacuerdo"  
## [3] "Ni de acuerdo ni en desacuerdo" "De acuerdo"  
## [5] "Totalmente de acuerdo"
```



## cut

La función `cut()` nos ayuda a convertir una variable numérica en factores. El parámetro `breaks` indica el número de intervalos, o los puntos de corte.

```
x <- runif(1000, 0, 10)
length(x)
```

```
## [1] 1000
```

```
f <- cut(x, breaks=4)
summary(f)
```

```
## (-0.00751,2.5]      (2.5,4.99]      (4.99,7.48]      (7.48,9.99]
##                251                250                241                258
```

```
f <- cut(x, breaks=c(0, 5, 7, 9, 10),  
        labels=c("SUSPENSO", "APROBADO",  
                  "NOTABLE", "SOBRESALIENTE"))  
  
summary(f)
```

##	SUSPENSO	APROBADO	NOTABLE	SOBRESALIENTE
##	503	204	188	105