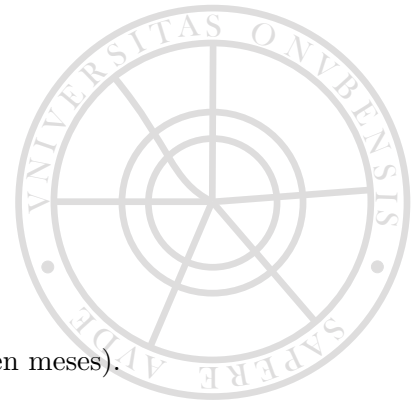
# Estadística. Práctica 3

# Estadística descriptiva de varias variables

En esta práctica veremos algunas herramientas descriptivas para estudiar conjuntamente varias variables y buscar relaciones entre ellas. Trabajaremos con el fichero **R_employee_data** que contiene información acerca de los empleados de una empresa sobre los que se observan las siguientes variables:

- *sexo*: refleja si el empleado es hombre o mujer.
- *educ*: nivel educativo.
- *catlab*: categoría laboral del empleado.
- *salario*: salario actual del empleado.
- *salini*: salario del empleado al ingresar en la empresa.
- *tiempemp*: antigüedad del empleado en la empresa (en meses).
- *expprev*: experiencia del empleado antes de ingresar en la empresa (en meses).




Nos va a interesar también en esta práctica calcular una nueva variable que indique la proporción, respecto del salario inicial, del aumento salarial de cada empleado. El nombre y expresión de esta variable es:

$$aumento = (salario - salini)/salini$$

## 3.1 Coeficiente de correlación lineal y gráfica XY

Para estudiar si dos variables numéricas están relacionadas podemos calcular el coeficiente de correlación lineal de Pearson y/o dibujar la nube de puntos asociada a ese par de variables.

Mediante `Estadísticos` → `Resúmenes` → `Matriz de correlaciones` podemos calcular el coeficiente de correlación entre los pares de variables numéricas que hay en el archivo.

La gráfica XY (o nube de puntos) complementa la información obtenida por el coeficiente de correlación. Se obtiene a través de `Gráficas` → `Gráfica XY`. La `variable explicativa` es la

que se coloca en el eje $X$, mientra la `variable explicada` se coloca en el eje $Y$. Si elegimos además una variable cualitativa en el campo `Condiciones` o en el campo `Grupos` obtenemos un gráfico XY de las variables $X$ e $Y$ para cada una de las modalidades de la variable elegida, con la diferencia de que en el primer caso los gráficos se representan en distintos ejes y en el segundo sobre los mismos ejes.

Observación: A través de `Gráficas` → `Diagrama de dispersión` obtenemos también gráficas XY con la posibilidad de dibujar la recta de regresión, entre otras opciones.

**Ejercicios:**

1. ¿Cuál es el coeficiente de correlación lineal de Pearson entre *salario* y *salini*? ¿y entre *expprev* y *tiempemp*? Representar la gráfica XY de cada una de estas parejas de variables.
2. Utilizar una gráfica XY para estudiar la relación entre el salario actual y la antigüedad en la empresa en cada categoría laboral.

## 3.2 Resúmenes numéricos por grupos

Como comentamos en la práctica anterior, mediante `Estadísticos` → `Resúmenes` → `Resúmenes numéricos` podemos calcular medidas descriptivas de una o varias variables numéricas, con la posibilidad de obtenerlas según las modalidades definidas por una variable cualitativa, que elegimos en `Resumir por grupos`.

**Ejercicio:** Calcular la media, la desviación típica, el coeficiente de variación, la asimetría, el apuntamiento y los cuartiles de las variables numéricas del archivo según la variable *catlab*.

## 3.3 Histogramas por grupos

Además de hacer resúmenes numéricos por grupos, para estudiar una variable numérica según las modalidades definidas por una variable cualitativa podemos representar y comparar los correspondientes histogramas de porcentajes.

**Ejercicio:** Representar los histogramas de porcentajes de la variable *expprev* según la variable *catlab*.

## 3.4 Tabla de estadísticas

Si queremos calcular medidas descriptivas de variables numéricas según una **o más** variables cualitativas o factores, elegimos `Estadísticos` → `Resúmenes` → `Tabla de estadísticas`.

**Ejercicios:**

1. Obtener el aumento mediano de salario los empleados según las variables *sexo* y *educ*.
2. Tras convertir la variable *tiempemp* en un factor de nombre *factor_tiempemp*, obtener el aumento medio de salario de los empleados según el sexo y la antigüedad en la empresa.

## 3.5 Gráfica de las medias

La última tabla obtenida es difícil de interpretar debido a la cantidad de valores que presenta la variable *factor_tiempemp*. Podemos obtener una representación gráfica de la información contenida en esta tabla que nos permita "visualizar" fácilmente dicha información. Para ello seleccionamos `Gráficas` → `Gráfica de las medias`. En `Factores` seleccionamos *factor_tiempemp* y *sexo*, y como `variable explicada` elegimos *aumento*. Por último, marcamos `sin barras de errores` y tenemos el gráfico que representa a la tabla del ejercicio 2 anterior.

## 3.6 Tabla de contingencia

Para el estudio conjunto de dos o más variables cualitativas se utiliza una **tabla de contingencia**.

Supongamos que estamos interesados en conocer cómo se reparten los empleados de esta empresa, no sólo según su categoría laboral (*catlab*), sino también según su nivel educativo (*educ*). Para ello seleccionamos en el menú `Estadísticos` → `Tablas de contingencia` → `Tabla de doble entrada`. En el espacio `Variable de fila` elegimos una de las dos variables, por ejemplo *educ* y en el espacio `Variable de columna` la otra, en nuestro caso, *catlab*. A continuación, seleccionamos la opción `Sin porcentajes` y desmarcamos la opción `Test de independencia Chi-cuadrado`. Si quisiéramos hacer el estudio para un subconjunto de empleados, lo indicaríamos en `Expresión de selección`, pero no es nuestro caso.

El número que aparece en cada celda representa la cantidad de empleados que verifican a la vez las condiciones dadas por su fila y su columna. Si lo que nos interesa es el porcentaje de empleados que corresponden a cada casilla, al realizar la tabla debemos marcar la opción `Porcentajes totales`.

Si en lugar de buscar cómo se distribuyen los empleados según ambas variables queremos ver, por ejemplo, cómo se distribuye el nivel Educativo dentro de cada una de las categorías laborales, el procedimiento es análogo, pero seleccionamos la opción de `Porcentajes por columnas`, ya que la variable *Categoría laboral* la pusimos en el espacio `Variable de columna`. Una vez hecho esto, obsérvese que los tantos por ciento por columnas suman 100% y que en la casilla *catlab: Administrativo*, y *educ: 8* aparece un 11%, lo que significa que el 11% de los administrativos tienen nivel educativo 8. De forma análoga, se puede estudiar cómo se distribuye la categoría laboral según el nivel educativo seleccionando en este caso `Porcentajes por filas`.

**Ejercicio:** A partir de las tablas construidas, responder a las siguientes preguntas:

(a) ¿Cuántos administrativos tienen nivel educativo menor que 15?

(b) ¿Cuántos empleados tienen nivel educativo menor que 15?

(c) ¿Qué porcentaje de empleados tienen nivel educativo 12 y son de seguridad?

(d) ¿Qué porcentaje de directivos tienen nivel educativo mayor que 19?

(e) ¿Qué porcentaje de empleados de nivel educativo 12 son administrativos?