Análisis estadístico de asociación entre variables

Autor: Roberto P. Muñoz
Cargo: Data Scientist en MetricArts
Github: https://github.com/rpmunoz

En la sección anterior aprendimos cómo hacer un análisis descriptivo usando una sola variable. Para ello empleamos elementos estadísticos tales como tablas de frecuencia, medidas de localización y medidas de dispersión.

En esta sección aprenderemos cómo hacer un análisis estadístico de la asociación entre variables. Entendemos como asociación el análisis de relaciones, covarianza y correlación entre variables.

El análisis estadístico de la asociación entre variables representa una parte fundamental del análisis de datos, pues muchas de las preguntas e hipótesis que se plantean en los estudios implican analizar la existencia de relación entre variables.

Análisis de asociación

La **existencia** de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables.

- Supongamos que tenemos una variable X que puede tomar los valores [a, b, c] y otra variable Y que puede tomar los valores [m, n, p]. Los datos empíricos indican que los sujetos que en X tienen valor **a** en Y tienden a tener valor **n**, que las que son **b** tienden a ser **p**, y que las que son **c** tienden a ser **m**. Esta evidencia pone de manifiesto que existe cierta asociación entre ambas variables.

Para visualizar esta situación, usamos las distribuciones de frecuencias de la variable Y para aquellos casos que en la variable X toma los valores **a**, **b** y **c**, respectivamente.

Complementariamente, se habla de **independencia** entre variables cuando no existe tal patrón de relación entre los valores de las mismas.

- Siguiendo el ejemplo anterior, sería el caso en que los sujetos que en X son **a**, en Y tienen una distribución que es igual o muy similar a la que tienen los que en X son **b** y **c**.

Para visualizar esta situación, usamos las distribuciones de frecuencias de la variable Y.

La asociación entre variables no debe entenderse como una cuestión de todo o nada, sino como un continuo que iría desde la ausencia de relación (independencia) al nivel

máximo de relación entre las variables. Este grado máximo se plasmaría en una relación determinista, esto es, el caso en que a partir del valor de un sujeto cualquiera en una variable, se puede afirmar cual será su valor en la otra variable.

Análisis usando Python

In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

%matplotlib inline
params = {'legend.fontsize': 'x-large',
          'figure.figsize': (10, 7.5),
         'axes.labelsize': 'x-large',
         'axes.titlesize':'x-large',
         'xtick.labelsize':'x-large',
         'ytick.labelsize':'x-large'}
plt.rcParams.update(params)

np.set_printoptions(precision=2)
%precision 2
```

Out[2]:

```
'%.2f'
```

Lectura de datos

En este tutorial usaremos una base de datos que contiene información con los precios de venta de 600 casas.

Los campos disponibles son la fecha en que fue vendida la casa, el precio en dólares americanos, el área útil en metros cuadrados, el área construida en metros cuadrados, el número de habitaciones, el número de baños, el número de pisos y el año en que fue construida.

In [ ]:

In [3]:

```
casas_file='data/precios_casas.csv'
casas=pd.read_csv(casas_file)
casas.head()
```

Out[3]:

	fecha	precio_USD	area_m2_util	area_m2_construida	habita
0	2014-05-29	485000.0	148.6	399.5	4
1	2015-03-04	570000.0	117.1	309.2	3
2	2014-06-26	518500.0	147.7	102.4	3
3	2014-06-13	822500.0	215.5	460.8	5
4	2014-11-04	511000.0	132.9	321.0	3

In [4]:

```
print("Número de casas: ", len(casas))
print("Número de habitaciones: ", sorted(casas['habitaciones'].unique()))
```

```
Número de casas:  602
```

```
Número de habitaciones:  [1, 2, 3, 4, 5, 6, 7, 9, 33]
```

En general, la manera más rapida de verificar si existe una relación entre variables es mediante el uso de visualizaciones. La visualización más empleada para este caso es el gráfico de puntos, el cual permite rápidamente ver la existencia de alguna relación.

Supongamos que queremos ver si existe alguna relación entre las variables **precio_USD** y **area_m2_util** del conjunto de datos **casas**. Para ello usamos la función plot.scatter() de la librería pandas.

```
In [5]:

casas.plot.scatter(x='area_m2_util', y='precio_USD', s=3);
plt.xlabel('Área en metros cuadrados')
plt.ylabel('Precio de casa (USD)')
plt.title('Análisis de relación entre precio y área')
```

```
Out[5]:

<matplotlib.text.Text at 0x105c03a90>
```

A continuación veremos cómo medir la asociación entre dos variables cuantitativas. Los índices más utilizados en estadística para analizar la intensidad o tamaño del efecto de

la relación lineal entre dos variables son la **covarianza** y el **coeficiente de correlación lineal**.

Covarianza

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal.

Dadas dos variables estadísticas **x** e **y** definiremos la covarianza S_{xy} como:

$S_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x}) (y_i - \overline{y})$

Donde n corresponde al número total de registros, $\overline{x}$ corresponde a la media de la variable **x** e $\overline{y}$ corresponde a la media de la variable **y**.

Desarrollando algebraicamente la fórmula de la covarianza se puede llegar a una fórmula que se considera más conveniente, la cual queda expresada como:

$S_{xy} = \frac{1}{n} \big(\sum_{i=1}^n x_i \cdot y_i \big) - \overline{x} \cdot \overline{y}$

Usaremos la funcion cov() de pandas para determinar la covarianza entre algunos campos del conjunto de datos casas. Partiremos analizando la covarianza entre los campos **precio_USD** y **area_m2_util**.

In [6]:

```
casas[['precio_USD','area_m2_util']].cov()
```

Out[6]:

	precio_USD	area_m2_util
precio_USD	4.366086e+10	9.069569e+06

	precio_USD	area_m2_util
area_m2_util	9.069569e+06	3.444799e+03

El resultado de la función cov() corresponde a una matriz que entrega la covarianza entre todas las variables de interés. El valor que nos interesa corresponde a la fila 1 y columna 2 de la matriz, la cual representa la covarianza entre los campos **precio_USD** y **area_m2_util**.

Podemos notar que el valor de la covarianza es 9.07×10^6. Este valor es positivo y mucho mayor que cero, por lo cual podemos concluir que existe una dependencia directa o positiva entre ambas variables.

Analicemos la covarianza entre las columnas **precio_USD** y **año_construida**.

In [7]:

```
casas[['precio_USD','año_construida']].cov()
```

Out[7]:

	precio_USD	año_construida
precio_USD	4.366086e+10	-1.213920e+06
año_construida	-1.213920e+06	1.725496e+03

Podemos notar que el valor de la covarianza es -1.21×10^6. Este valor es negativo y mucho menor que cero, por lo cual podemos concluir que existe una dependencia inversa o negativa entre ambas variables.

Coeficiente de correlación

Si bien la covarianza cumple el objetivo de medir el grado de asociación entre dos variables, ésta sufre de algunos inconvenientes. El primer inconveniente es que no tienen valores máximo ni mínimo, y el segundo es que su valor depende de las unidades de medida de las variables. Dada esta situación, es conveniente definir la correlación y sus respectivos coeficientes.

La correlación es una cantidad que indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas.

Existen diversos coeficientes que miden el grado de correlación, adaptados a la naturaleza de los datos. El más conocido es el coeficiente de correlación de Pearson, que se obtiene dividiendo la covarianza de dos variables entre el producto de sus desviaciones estándar.

Dadas dos variables estadísticas **x** e **y**, el coeficiente de correlación de Pearson r_{xy} se define como:

$r_{xy} = \frac{S_{xy}}{S_x S_y}$

El coeficiente de correlación de Pearson se interpreta de modo análogo a la covarianza pero, al oscilar entre -1 y 1 como máximo, la interpretación del mismo resulta más intuitiva a la vez que facilita el establecimiento de comparaciones entre los coeficientes obtenidos para conjuntos de datos distintos.

Usaremos la funcion corr() de pandas para determinar el coeficiente de correlatión entre algunos campos del conjunto de datos **casas.** Partiremos analizando la correlación entre los campos **precio_USD** y **area_m2_util**.

In [8]:

```
casas[['precio_USD','area_m2_util']].corr()
```

Out[8]:

	precio_USD	area_m2_util
precio_USD	1.000000	0.739535
area_m2_util	0.739535	1.000000

El resultado de la función corr() corresponde a una matriz que entrega el coeficiente de correlación entre todas las variables de interés. El valor que nos interesa corresponde a la fila 1 y columna 2 de la matriz, la cual representa la correlación entre los campos **precio_USD** y **area_m2_util**.

Podemos notar que el valor del coeficiente de correlación es 0.74. Este valor es positivo y cercano a uno, por lo cual podemos concluir que existe una relación directa entre ambas variables y su correlación es muy alta.

Analicemos la covarianza entre las columnas **precio_USD** y **año_construida.**

In [9]:

```
casas[['precio_USD','año_construida']].corr()
```

Out[9]:

	precio_USD	año_construida
precio_USD	1.000000	-0.139858

	precio_USD	año_construida
año_construida	-0.139858	1.000000

Podemos notar que el valor del coeficiente de correlación es -0.14. Este valor es negativo y más cercano al valor cero, por lo cual podemos concluir que existe una dependencia inversa entre ambas variables y su correlación es baja.

```
In [ ]:
```

```
In [ ]:
```

Content source:
MetricLearning/diplomado_datascience

Similar notebooks:

- 02_Analisis_de_correlacion
- simulation_thesis_comp
- Comparación entre dos emisoras
- Regresión Lineal
- QSAR
- quantitativePython
- Lycees_2015_BAC
- Modulo3
- GyC - Practica 1 - R

- 1. Similaridade por cosseno-checkpoint