Análisis de Datos con Python
Nota de Clase 01: Estimados de Locación y Variabilidad

Manuel Soto Romero

17 de septiembre de 2020
BEDU Tech

[18]:
```
import pandas as pd
import numpy as np
```

1 Objetivos

- Utilizar Google Colab en conjunción con Google Drive y Github.
- Identificar los tipos de datos estructurados existen.
- Identificar los estimados de locación y en qué momento son útiles.
- Identificar valores típicos y atípicos.
- Realizar cálculos estadísticos robustos.
- Identificar los estimados de variabilidad y en qué momento son útiles.
- Identificar los estadísticos de orden.

2 Contenido

2.1 Utilización de software

Antes que nada, vamos a asegurarnos de que las siguientes cosas estén resueltas:

a) Todos tenemos una cuenta en Google Drive.
b) Todos hemos creado un Acceso Directo desde la carpeta Datasets a nuestro Drive.
c) Todos sabemos cómo montar nuestro Google Drive en un Notebook de Google Colab y leer nuestros conjuntos de datos usando `pandas`.
d) Todos sabemos cómo leer Jupyter Notebooks desde el repositorio del módulo en Google Colab.

2.2 Datos estructurados

Hay dos grandes categorías de datos que manejamos en Ciencia de Datos: datos estructurados y datos no estructurados. Por el momento sólo vamos a hablar sobre datos estructurados. Los datos estructurados pueden subdividirse de la siguiente manera:

1. Numéricos: Datos representados por números que pueden tomar una cantidad no predefinida de valores.

a) Discretos: Datos que sólo pueden tomar el valor de un número entero.

b) Continuos: Datos que pueden toman cualquier valor dentro de un intervalo.

[24]:
```
# Datos numéricos
datos = pd.DataFrame({'mes': [1,2,3,4], 'temperatura': [28.5,10.2,36.5,40.87]})
datos
```

[24]:
```
   mes  temperatura
0    1        28.50
1    2        10.20
2    3        36.50
3    4        40.87
```

[25]:
```
# El mes es discreto
datos.mes.unique()
```

[25]:
```
array([1, 2, 3, 4], dtype=int64)
```

[26]:
```
# La temperatura es continua
datos.temperatura.unique()
```

[26]:
```
array([28.5 , 10.2 , 36.5 , 40.87])
```

2. Categóricos: Datos que sólo pueden tomar un conjunto específico de valores que representan un conjunto de posibles categorías.

a) Binarios: Datos categóricos que sólo tienen dos categorías posibles.
b) Ordinales: Datos categóricos que tienen un orden explícito.

Los datos categóricos los examinamos bien con tablas de frecuencias o con representaciones gráficas como diagramas de barras o de pastel.

[12]:
```
# Datos categóricos
datos = pd.read_csv("ejemplo.csv")
datos
```

[12]:
```
     sexo nivelest tabaco estcivil laboro  hijos  edad peso talla  \
0       2        1      2        3      4      2    67   72   159
1       1        4      0        2      3      1    56  150   178
2       1        2              4      3      0    81   71   158
3       2        1      0        2      4      2    74   85   188
4       1        5      2        2      1      3    53  102   178
..    ...      ...    ...      ...    ...    ...   ...  ...   ...
531     2        2      0        4      4      3    69
532     1        2      0        2      3      2    70
533     1        3      3        2      1      2    42
534     1        3      2        2      1      4    53        175
535     1        2              2      3      3    69        165

                imc sedentar diabm hipercol  pas  pad  fc
```

```
0                   29        1     2        2  190  100   9
1    47.342507259184444       1     2        2  192   79  53
2                   30        1     2        1  190   95   9
3    24.049343594386603       1     1        1  193   90  82
4     32.19290493624542       1     1        1  182  114  73
..                  ...     ...   ...      ...  ...  ...  ..
531                              1        2  151   88  71
532                       1      2        2  160   73  83
533                       1      2        2  157  108  98
534                              2        2  120   80  68
535                       1      1        1  130   80  56

[536 rows x 16 columns]
```

[15]:
```
# El sexo es binario
datos.sexo.unique()
```

[15]:
```
array([2, 1], dtype=int64)
```

[16]:
```
# El nivel de estudios es ordinal
datos.nivelest.unique()
```

[16]:
```
array(['1', '4', '2', '5', '3', ' '], dtype=object)
```

Ir al reto 1

2.3 Estimados de locación

Los estimados de locación nos sirven para determinar qué valor describe mejor un conjunto de datos. A este valor le llamamos el "valor típico" de nuestro conjunto. Dos estimados son los más comunes y lo más utilizados:

1. Promedio (o media): es la suma de todos los valores que conforman una muestra, divididos en la cantidad total de datos.
2. Mediana: es el valor medio de los dos valores que se encuentran justo en la mitad de un grupo de datos. Si se tiene un número par de datos, se deben sumar los dos valores centrales y dividir su resultado entre 2, así se obtiene la mediana. Si el número de datos es impar, el dato del centro es la mediana de la muestra.

Veamos cómo se calculan usando `pandas`.

Ir al ejemplo 1
Ir al reto 2

2.4 Valores atípicos

Hemos aprendido a conseguir "valores típicos" se sirven para describir un conjunto de datos. Así como existen valores que se parecen a la norma, hay también valores que difieren mucho de

la norma. Estos valores suelen sobresalir de nuestro conjunto de datos porque se encuentran a mucha distancia del grueso de las muestras.

Estos valores son los "valores atípicos". Hay veces en las que tener valores atípicos no nos preocupa, pero otras veces pueden ser peligrosos porque modifican nuestros estimados de locación, dándonos descripciones erróneas de nuestro conjunto de datos.

Más adelante aprenderemos a detectar valores atípicos, pero por lo pronto vamos a aprender un estimado de locación que es más *robusto* (menos sensible a valores atípicos) que el promedio y la mediana: la media truncada.

Ir al ejemplo 2

2.5 Estimados de variabilidad

Los estimados de locación se utilizan para intentar encontrar un "valor típico" que describa adecuadamente nuestro conjunto de datos. Los estimados de variabilidad, en cambio, nos sirven para determinar qué tan dispersos están los datos alrededor de nuestro valor típico. Nuestros datos pueden estar en general o muy cerca o muy distantes del valor típico, y los estimados de variabilidad nos ayudan a determinar esto.

Uno de los estimados más utilizados es la desviación estándar. Veamos cómo funciona.

Ir al ejemplo 3
Ir al reto 3

2.6 Estadísticos de Orden

Existen otras formas de calcular la dispersión de nuestros datos que requieren que nuestros datos estén ordenados ascendentemente. Estos cálculos nos dan otra perspectiva acerca de la distribución de nuestros datos que puede sernos de mucha utilidad.

Tres de los estadísticos de orden más comunes son el **Rango**, los **Percentiles** y el **Rango intercuartílico**. Veamos cómo funcionan.

Ir al ejemplo 4
Ir al reto 4

3 Postwork

`Postwork Sesión 1`