Análisis de Datos: Dispersión de variables numéricas

Para saber si los datos están muy dispersos o se concentran en torno a un valor (caso en el que una de las medidas centrales como la media o la mediana los representarán muy bien), hemos visto ya los percentiles y el rango para hacernos una idea, pero en general lo que se emplea en el caso de los valores numéricos son dos medidas, relacionadas entre sí, y luego métodos visuales basados en dos tipos de gráficas: los histogramas y las gráficas de función de densidad de probabilidad [Aunque a esta parte le dedicaremos la siguiente píldora]

Vamos a verlos y a aplicarlos a nuestros dos casos de uso de ejemplo.

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

df_seguros = pd.read_csv("./data/Marketing-Customer-Analysis.csv")
df_air_jun = pd.read_csv("./data/dataset_viajes_jun23.csv")
```

Medidas de dispersión o variabilidad: Varianza

La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de un conjunto de datos (por ejemplo, los valores de una de nuestras variables o columnas y, en general, una distribución estadística). La varianza intenta describir la dispersión de los *datos*. *Básicamente representa lo que varían los datos*. **Como está elevada al cuadrado, la varianza no puede tener las mismas unidades que los datos**.

Una varianza elevada significa que los datos están más dispersos. Mientras que un valor bajo, indica que los datos están próximos a la media.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

INCISO: Y por qué está elevado al cuadrado te preguntarás: Porque no quiero que las diferencias positivas y negativas se compensen. Piensa en esta serie de datos:

[-104,100,102,0,-100,120,-119]

La media es:

$$\mu = \frac{1}{n}\sum_i x_i = \frac{(-104 + 100 + 102 + 0 - 100 + 120 - 119)}{7} = \frac{-1}{7} \approx -0.14$$

Si no eleváramos al cuadrado, el denominador sería:

$$\sum_{i=1}^{n}(x_i - \mu) = (-104 - 0.14) + (100 - 0.14) + (102 - 0.14) + (0 - 0.14) + (-100 - 0.14) + (120 - 0.14) + (-119 - 0.14) = -1.98$$

Y al dividirlo se nos quedaría en un grado de dispersión de 0.28 que no es real (fíjate que el rango como tal es 239), para que no se compensen las diferencias se elevan al cuadrado.

Y la varianza según la definición de ese conjunto de datos es:

$$\sigma^2 \approx 9968.69$$

En vez de aplicarla a los datos veamos la versión comparable (es decir medida en las mismas unidades que los datos que estamos analizando) que es la desviación estándar.

Medidas de dispersión o variabilidad: Desviación estándar

La desviación típica es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Una ventaja que presenta la desviación estándar sobre la varianza es que se expresa en unidades de la variable en cuestión.

¿Y para qué nos sirven estas medidas?

- Nos dan una idea rápida de si los datos están dispersos (gte. compararemos la desviación con la media) y por tanto de si la media y mediana son buenos representantes de los valores o bien tenemos que trabajar la variable de otra forma, considerando rangos por ejemplo.
- Nos sirven para hacer cálculos posteriores y de otras medidas (que nos permitirán interpretar los datos de otras formas)

Podemos aplicar directamente el concepto de "Coeficiente de Variación" (CV) que es la división de la desviación estándar entre la media. Como pautas generales:

- Un CV menor al 15% suele considerarse como una baja variabilidad.
- Un CV entre 15% y 30% indica una variabilidad moderada.
- Un CV mayor al 30% a menudo se considera como una alta variabilidad.

Estos valores son orientativos y deben interpretarse en el contexto específico de tus datos y el área de estudio.

Caso 1. Seguros: Dispersión

En general fijate en la desviación y de nuevo la podemos obtener del método describe():

df_seguros.describe()

[2]:

	customer_lifetime_value	income	monthly_premium_auto	months_since_last_claim	months_since_policy_inception	number_of_open_complaints	number_of_policies
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000
mean	8004.940475	37657.380009	93.219291	15.097000	48.064594	0.384388	2.966170
std	6870.967608	30379.904734	34.407967	10.073257	27.905991	0.910384	2.390182
min	1898.007675	0.000000	61.000000	0.000000	0.000000	0.000000	1.000000
25%	3994.251794	0.000000	68.000000	6.000000	24.000000	0.000000	1.000000
50%	5780.182197	33889.500000	83.000000	14.000000	48.000000	0.000000	2.000000
75%	8962.167041	62320.000000	109.000000	23.000000	71.000000	0.000000	4.000000
max	83325.381190	99981.000000	298.000000	35.000000	99.000000	5.000000	9.000000

Si obtenemos el coeficiente de variación:

```
def variabilidad(df):
    df_var=df.describe().loc[["std","mean"]].T
    df_var["CV"]=df_var["std"]/df_var["mean"]
    return df_var
```

variabilidad(df_seguros)

[8]:

	std	mean	CV
customer_lifetime_value	6870.967608	8004.940475	0.858341
income	30379.904734	37657.380009	0.806745
monthly_premium_auto	34.407967	93.219291	0.369108
months_since_last_claim	10.073257	15.097000	0.667236
months_since_policy_inception	27.905991	48.064594	0.580594
number_of_open_complaints	0.910384	0.384388	2.368397
number_of_policies	2.390182	2.966170	0.805814
total_claim_amount	290.500092	434.088794	0.669218

Conclusión (no para mostrar sino para seguir avanzando en el EDA):

- Salvo quizá monthly_premium_alto para el resto de las variables debo analizar con cuidado esa distribución de valores.
- En el caso de CLV que es una de nuestras directoras, y en el de income que es importante, tendré que dar más cariño

Caso 2. Viajes: Dispersión

Aplicando ya directamente la función:

variabilidad(df_air_jun)

[9]:

	std	mean	CV
distancia	5550.244086	8071.003333	0.687677
consumo_kg	67441.849592	68240.520508	0.988296
duracion	450.474786	635.873333	0.708435
ingresos	318285.763970	418768.851500	0.760051

Nos ocurre algo parecido, en este caso probablemente sería más interesante hacerlo por compañía todo, pero eso lo veremos en la siguiente unidad. En cualquier caso, tanto consumo (CV > 90%) como ingresos, nuestras variables directoras (o por lo menos las que nosotros hemos decidido que lo sean por ahora) necesitan que miremos sus distribuciones.

En ambos casos tenemos que ver qué "pinta" tienen los datos, no nos vale con estas pistas de las métricas y ahí es donde entran nuestras dos siguientes herramientas: Los histogramas y las funciones de densidad.