Práctica 4

Estadística descriptiva unidimensional: transformaciones lineales

En esta práctica se describe cómo realizar transformaciones lineales y, en particular, cómo tipificar variables. Además, se abordan otros aspectos como el efecto de las transformaciones lineales sobre algunas de las medidas de síntesis estudiadas.

Contenido de la práctica

Ejercicio 28: Un céntrico hotel zaragozano ha recibido 1690 valoraciones a través de internet, donde los usuarios han mostrado su opinión utilizando una escala de 1 a 5 estrellas (☆☆☆☆☆). El archivo *valoraciones.txt* contiene el número de estrellas que ha proporcionado cada usuario. Impórtalo y realiza un análisis descriptivo de la variable, mediante tablas y gráficos. Después, calcula la valoración media de los usuarios y la desviación típica.

4.1. Transformaciones lineales

Vamos a repasar las transformaciones lineales y sus propiedades resolviendo el siguiente ejercicio:

Ejercicio 29: Realiza los siguientes apartados:

1. Crea una variable llamada *valoracion0a4*, que contenga la valoración de los usuarios en una escala de 0 a 4.
2. A partir de la variable *valoracion0a4*, crea un variable llamada *valoracion0a10*, que contenga la valoración de los usuarios en una escala de 0 a 10.
3. ¿Cuál es la transformación lineal que se le ha aplicado a la variable *valoración* para obtener la variable *valoracion0a10*?

4. Calcula, utilizando R Commander como calculadora, cuál debería ser la media de la variable *valoracion0a10*.
5. Calcula la media de la variable *valoracion0a10* y comprueba que el resultado coincide con el cálculo realizado en el apartado anterior.
6. Calcula, utilizando R Commander como calculadora, cuál debería ser la desviación típica de la variable *valoracion0a10*.
7. Calcula la desviación típica de la variable *valoracion0a10* y comprueba que el resultado coincide con el cálculo realizado en el apartado anterior.
8. Calcula la tabla de frecuencias de la variable *valoracion0a10* y compárala con la de la variable *valoracion*.

4.1.1. Tipificación de variables

Para tipificar los valores que toma una variable X, se le resta su media y el resultado se divide por su desviación típica:

$$Z = \frac{X - \bar{x}}{s_x}$$

Ejercicio 30: Realiza los siguientes apartados:

1. Tipifica las variables *valoracion* y *valoracion0a10* y llámalas *T.valoracion* y *T.valoracion0a10*. Utiliza para ello la opción:

   Datos > Modificar variables del conjunto de datos > Calcular una nueva variable...

2. Comprueba que las medias y desviaciones típicas de las variables tipificadas son 0 y 1, respectivamente.
3. Visualiza y compara las variables tipificadas.
4. Calcula la tabla de frecuencias de la variable *T.valoracion* y compárala con la de la variable *valoracion*.

R Commander tiene una opción específica para tipificar variables sin necesidad de calcular previamente la media y la desviación típica:

Datos > Modificar variables del conjunto de datos activo > Tipificar variables...

Este menú no permite asignar el nombre de la nueva variable, aunque es posible cambiarlo pulsando el botón *Editar conjunto de datos*.

Ejercicio 31: Tipifica las variables *valoracion* y *valoracion0a10* utilizando la opción específica para ello. Después visualiza las variables tipificadas y compáralas con el resto.

Nota: en muchas aplicaciones informáticas, debido a la precisión en el cálculo, hay números hemos de interpretar como cero. Por ejemplo, un número muy cercano a cero puede tener varias formas:

$$0.0000000001234567 = \texttt{1.234567e-10} = 1.234567 \cdot 10^{-10}$$

Todas ellas representan el mismo número que claramente puede interpretarse como cero. Observa que lo importante es que después de la letra `e` haya un número negativo que es el que nos indica los decimales con valor cero que tendremos.

4.2. Ejercicios propuestos

Ejercicio 32: El archivo *Encuesta Continua Hogares INE 2020.xlsx* contiene algunos de los datos obtenidos por el Instituto Nacional de Estadística en la Encuesta Continua de Hogares del año 2020. En particular, la variable *SuperficieVivienda* indica la superficie útil de cada una de las viviendas analizadas. En primer lugar:

- Filtra el conjunto de datos (sin cambiarle el nombre) y guarda únicamente aquellos datos de viviendas con una superficie de $300\,\mathrm{m}^2$ o menos.
- Tipifica la variable *SuperficieVivienda* y llámala *Z.SuperficieVivienda*.

Ahora, responde las siguientes preguntas que hacen referencia a la variable tipificada *Z.SuperficieVivienda*:

1. Compara la media y la desviación típica de las variables *SuperficieVivienda* y *Z.SuperficieVivienda*.
2. Compara los histogramas de las variables *SuperficieVivienda* y *Z.SuperficieVivienda*.
3. ¿Cuál son las viviendas cuya superficie está más alejada de la media, tanto por encima como por debajo de ella? ¿Cuántas desviaciones típicas se alejan en cada caso?
4. Calcula una nueva variable denomina *nDesv*, de tipo factor, que tome:
   - El valor 1 si la variable tipificada está en $[-1, 1]$.
   - El valor 2 si la variable tipificada está en $[-2, -1)$ o en $(1, 2]$.
   - El valor 3 si la variable tipificada está en $[-3, -2)$ o en $(2, 3]$.
   - etc.

   ¿Qué representa la variable *nDesv*?
5. Calcula, haciendo uso de la tabla de frecuencias de la variable *nDesv*, cuántos datos distan de la media:
   - 1 desviación típica o menos.
   - 2 desviaciones típicas o menos.
   - 3 desviaciones típicas o menos.
   - 4 desviaciones típicas o menos.
6. Comprueba que se cumple la desigualdad de Chebychev. Recuerda que dicha desigualdad nos dice que al menos el $100(1 - 1/k^2)\,\%$ de los datos están a una distancia de la media de, como mucho, k desviaciones típicas.
7. ¿Estamos ante una distribución Normal (aproximadamente)? Comprueba en qué medida se verifican los siguientes aspectos:
   - La mayor parte de los datos están cerca de la media.
   - La distribución es simétrica.
   - Media, mediana y moda coinciden.
   - El coeficiente de apuntamiento es 0.
   - Se cumple la regla empírica: 68 % - 95 % - 99.7 %.

Ejercicio 33: Introduce manualmente la variable *Celsius*, correspondiente a la temperatura máxima registrada en junio en grados Celsius en ciertas ciudades españolas, según la tabla siguiente:

Ciudad	**X(°C)**
Almería	40.8
Barcelona	34.9
Bilbao	41.2
Granada	41.5
Madrid	40.0
Murcia	42.5
Oviedo	35.5
Sevilla	45.2
Teruel	38.0
Zaragoza	41.0

Después, calcula las variables:

- *Fahrenheit*: temperatura en grados Fahrenheit, 1.8·*Celsius*+32.
- *CelsiusCent*: variable *Celsius* centrada.
- *CelsiusReesc*: variable *Celsius* reescalada.
- *CelsiusTipi*: variable *Celsius* tipificada.
- *FahrenheitTipi*: variable *Fahrenheit* tipificada.

y guarda el fichero obtenido en el fichero *Temperatura.Rdata*.

Calcula la media y la desviación típica de todas las variables y comprueba que toman los valores esperados.