Taller 1 - Programación en Lenguajes Estadísticos

Manuel Rodríguez
manrodriguezar@unal.edu.co
Cristian Padilla
cpadilla@unal.edu.co
Triana Ramírez
trramirezf@unal.edu.co

Agosto 2022

1. Traducción de la sección "Elements of structured data" (pags. 2-4) del libro "Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media".

Elementos de datos estructurados

Los datos provienen de muchas fuentes: mediciones de sensores, eventos, texto, imágenes y vídeos. El Internet de las cosas (IoT) está arrojando flujos de información. Gran parte de estos datos no están estructurados: las imágenes son una colección de píxeles, y cada píxel contiene RGB (rojo, verde, azul) información de color. Los textos son secuencias de palabras y caracteres que no son palabras, a menudo organizados por secciones, subsecciones, etc. Los flujos de clics son secuencias de acciones realizadas por un usuario que interactúa con una aplicación o una página web. De hecho, un gran desafío de la ciencia de datos es convertir este torrente de datos sin procesar en información procesable. Para aplicar los conceptos estadísticos que se tratan en este libro, los datos brutos no estructurados deben ser procesados y manipulados en una forma estructurada. Una de las formas más comunes de datos estructurados es una tabla con filas y columnas, ya que los datos pueden surgir de una base de datos relacional o recopilarse para un estudio. Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Los datos numéricos se presentan en dos formas: continuos, como la velocidad del viento o la duración del tiempo, y discretos, como el recuento de la ocurrencia de un evento. Los datos categóricos toman solo un conjunto

fijo de valores, como un tipo de pantalla de TV (plasma, LCD, LED, etc.) o el nombre de un estado (Alabama, Alaska, etc.). Los datos binarios son un caso especial importante de datos categóricos que toman solo uno de dos valores, como 0/1, sí/no o verdadero/falso. Otro tipo útil de datos categóricos son los datos ordinales en los que se ordenan las categorías; un ejemplo de esto es una calificación numérica (1, 2, 3, 4 o 5). ¿Por qué nos molestamos con la taxonomía de tipos de datos? Resulta que a los efectos del análisis de datos y el modelado predictivo, el tipo de datos es importante para ayudar a determinar el tipo de visualización, análisis de datos o modelo estadístico. De hecho, los software de ciencia de datos, como R y Python, utilizan estos tipos de datos para mejorar el rendimiento computacional. Más importante aún, el tipo de datos para una variable determina cómo el software manejará los cálculos para esa variable.

1.1. Términos claves para tipos de datos

- **Numérico**: datos que se expresan en una escala numérica.
- **Continuo**: datos que pueden tomar cualquier valor en un intervalo. (Sinónimos: intervalo, flotante, numérico).
- **Discreto**: datos que solo pueden tomar valores enteros, como recuentos. (Sinónimos: número entero, cuenta).
- **Categórico**: datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de categorías posibles. (Sinónimos: enumeraciones, enumerado, factores, nominal).
- **Binario**: un caso especial de datos categóricos con solo dos categorías de valores, por ejemplo, 0/1, verdadero/falso. (Sinónimos: dicotómico, lógico, indicador, booleano).
- **Ordinal**: datos categóricos que tienen un ordenamiento explícito. (Sinónimo: factor ordenado).

Los ingenieros de software y los programadores de bases de datos pueden preguntarse por qué necesitamos la noción de datos categóricos y ordinales para el análisis. Después de todo, las categorías son simplemente una colección de valores de texto (o numéricos), y la base de datos subyacente maneja automáticamente la representación interna. Sin embargo, la identificación explícita de los datos como categóricos, a diferencia del texto, ofrece algunas ventajas:

- Saber que los datos son categóricos puede actuar como una señal que le dice al software cómo deben comportarse los procedimientos, como producir un gráfico o ajustar un modelo. En particular, los datos ordinales se pueden representar como un factor ordenado en R, conservando un orden especificado por el usuario en gráficos, tablas y modelos. En Python, scikit learn admite datos ordinales con sklearn.preprocessing.OrdinalEncoder.

- El almacenamiento y la indexación se pueden optimizar (como en una base de datos relacional).
- Los valores posibles que puede tomar una variable categórica determinada se imponen en el software (como una enumeración).

El tercer "beneficio"puede dar lugar a un comportamiento no deseado o inesperado: el comportamiento predeterminado de las funciones de importación de datos en R (por ejemplo, read.csv) es convertir automáticamente una columna de texto en un factor. Las operaciones subsiguientes en esa columna supondrán que los únicos valores permitidos para esa columna son los que se importaron originalmente, y la asignación de un nuevo valor de texto introducirá una advertencia y producirá un NA (valor faltante). El paquete pandas en Python no realizará dicha conversión automáticamente. Sin embargo, puede especificar una columna como categórica explícitamente en la función read_csv.

1.2. Ideas claves

- Los datos normalmente se clasifican en el software por tipo.
- Los tipos de datos incluyen numéricos (continuos, discretos) y categóricos (binarios, ordinales).
- La tipificación de datos en el software actúa como una señal para el software sobre cómo procesar los datos.

1.3. Otras lecturas

- La documentación de pandas describe los diferentes tipos de datos y cómo se pueden manipular en Python.
- Los tipos de datos pueden resultar confusos, ya que los tipos pueden superponerse y la taxonomía de un software puede diferir de la de otro. El sitio web R Tutorial cubre la taxonomía de R. La documentación de pandas describe los diferentes tipos de datos y cómo se pueden manipular en Python.
- Las bases de datos son más detalladas en su clasificación de tipos de datos, incorporando consideraciones de niveles de precisión, campos de longitud fija o variable, y más; consulte la guía de SQL de W3Schools.

1.4. Datos rectangulares

El marco de referencia típico para un análisis en ciencia de datos es un objeto de datos rectangular, como una hoja de cálculo o una tabla de base de datos.

Datos rectangulares es el término general para una matriz bidimensional con filas que indican registros (casos) y columnas que indican características (variables); El marco de datos es el formato específico en R y Python. Los datos no siempre comienzan de esta forma: los datos no estructurados (p. ej., texto) deben procesarse y manipularse para que puedan representarse como un conjunto de características en los datos rectangulares (consulte "Elementos de los datos estructurados" en la página 2). . Los datos de las bases de datos relacionales deben extraerse y colocarse en una sola tabla para la mayoría de las tareas de modelado y análisis de datos.

2. Definiciones de "Medidas de tendencia central y dispersión": 1.Medidas de tendencia central (media aritmética, mediana y cuantiles, gráficos cuantil-cuantil, moda, media geométrica y media harmónica). 2.Medidas de dispersión (rango y rango intercuartil, desviación absoluta, varianza y desviación estándar, y coeficiente de variación).3.Diagramas de caja. 4.Medidas de concentración (curva de Lorenz y coeficiente Gini).

Medidas de tendencia central

Media: Media aritmética, es la que se obtiene sumando los datos y dividiéndolos por el número de ellos. Se aplica por ejemplo para resumir el número de pacientes promedio que se atiende en un turno. Otro ejemplo, es el número promedio de controles prenatales que tiene una gestante.

Media Geométrica (MG): Es una medida de tendencia central que puede utilizarse para mostrar los cambios porcentuales en una serie de números positivos. Como tal, tiene una amplia aplicación en los negocios y en la economía, debido a que con frecuencia se está interesado en establecer el cambio porcentual en las ventas en el producto interno bruto o en cualquier serie económica. Se define como la raíz índice n del producto de n términos.

Media armónica (Ma) La media armónica se define como el recíproco de la media aritmética de los recíprocos:

$$MG = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n}$$

Así por ejemplo, la media geométrica del conjunto de datos 2, 4, 6, 12 y 18 es

$$MG = \sqrt[5]{(2)(4)(6)(12)(18)}$$

$$MG = \sqrt[5]{10368}$$

$$MG = 6.355$$

$$Ma = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$$

Esta medida se emplea para promediar variaciones con respecto al tiempo tales como productividades, tiempos, rendimientos, cambios, etc.

Mediana: Es el valor que queda en el centro de los datos, una vez que estos sean ordenados en forma ascendente o descendente. Para hallar la mediana de un conjunto de datos se organiza de forma progresiva; si el conjunto de datos contiene un número impar de elementos el elemento de en medio del arreglo es la mediana; si hay un número par de observaciones la mediana es el promedio aritmético de los elementos centrales.

Ejemplo: Hallar la mediana de los siguientes datos que muestra N el número de pacientes tratados en la sala de emergencias durante 8 días consecutivos; 52, 35, 43, 11, 30, 31, 86 y 49.

Ordenamos los datos: 11, 30, 31, **35**, **43**, 49, 52, 86 (35+43)/2= 78/2= 39

Cuantiles (Ci) Los cuantiles son estadísticos de localización que sirven para estudiar o analizar lo que sucede con algún porcentaje de datos en particular, cuando se han ordenado previamente los datos. Los cuantiles se dividen en Cuartiles, deciles y percentiles; y se calculan teniendo en cuenta también, las tres formas en que se presentan los datos.

Cuartiles (Qi) Son aquellos números que dividen a éstas en cuatro partes porcentualmente iguales. Hay tres cuartiles, Q1, Q2 y Q3. El primer cuartil Q1, es el valor en el cual o por debajo del cual queda aproximadamente un cuarto (25 %) de todos los valores de la sucesión (ordenada); El segundo cuartil Q2 es el valor por debajo del cual queda el 50 % de los datos (Mediana), el tercer cuartil Q3 es el valor por debajo del cual quedan las tres cuartas partes (75 %) de los datos.

Deciles (Di) Los deciles se tienen cuando el conjunto de datos se divide en 10

partes iguales, de esta manera cada una de ellas acumula un 10 % del conjunto de datos. Por ejemplo, en D1 se acumula el 10 % de los datos, en D4 el 40 % y D8 se acumula el 80 %.

Percentiles (Pi) Los percentiles se tienen cuando el conjunto de datos se divide en 100 partes iguales, de esta manera cada una de ellas acumula un 1 % del conjunto de datos. Por ejemplo, en P1 se acumula el 1 % de los datos, en P45 el 45 % y P68 se acumula el 68 %.

Moda: Valor o (valores) que aparece(n) con mayor frecuencia o repetición. Ejemplo: Hallar la moda del siguiente conjunto: (2, 3, 3, 5, 3, 6, 9, 8, 5) = **3**

Gráficos Cuantil-Cuantil: Un gráfico Cuantil-Cuantil permite observar cuan cerca está la distribución de un conjunto de datos a alguna distribución ideal o comparar la distribución de dos conjuntos de datos.

Medidas de dispersión

Rango: El rango (R) o recorrido estadístico es la diferencia entre el valor máximo y el mínimo de un conjunto de elementos. **Fórmula del Rango**: R = (Max) – (Min)

Rango intercuartílico: El rango intercuartílico (IQR) (o rango intercuartil) es una estimación estadística de la dispersión de una distribución de datos. Consiste en la diferencia entre el tercer y el primer cuartil. Mediante esta medida se eliminan los valores extremadamente alejados. El rango intercuartílico es altamente recomendable cuando la medida de tendencia central utilizada es la mediana (ya que este estadístico es insensible a posibles irregularidades en los extremos).

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Varianza: La varianza (S2) mide la dispersión de los datos de una muestra respecto a la media, calculando la media de los cuadrados de las distancias de todos los datos.

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$ la media

Desviación típica o estándar: es una medida de dispersión (S) asociada a la media. Como estadístico, es la raíz cuadrada de la varianza. Es la raíz cuadrada del cuadrado de las desviaciones de los datos de una muestra (X1, X2,. . . , XN) de la media (x) dividido en el caso de la muestra por N-1. Está en las mismas unidades de los datos. Es un indicador de cómo tienden a estar agrupados los datos respecto a la media.

$$S_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{N-1}}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos

Coeficiente de Variación: El coeficiente de variación de Pearson (r) mide la variación de los datos respecto a la media, sin tener en cuenta las unidades en la que están.

$$r = \frac{S_X}{|\overline{x}|}$$

siendo S_X la desviación típica y $\overline{x}$ la media del conjunto de observaciones $(X_1, X_2, ..., X_N)$ y $\overline{x} \neq 0$

El coeficiente de variación toma valores entre 0 y 1. Si el coeficiente es próximo al 0, significa que existe poca variabilidad en los datos y es una muestra muy compacta. En cambio, si tienden a 1 es una muestra muy dispersa y la media pierde confiabilidad. De hecho, cuando el coeficiente de variación supera el 30 % (0,3) se dice que la media es poco representativa.

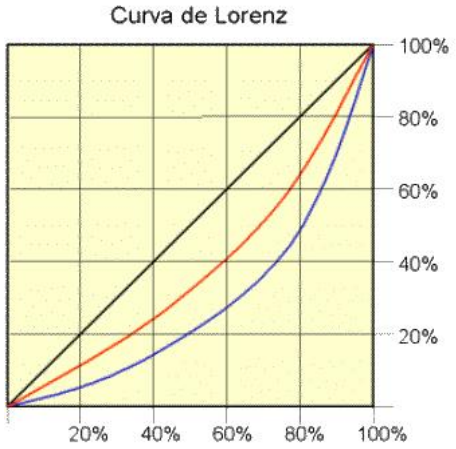
Desviación Media o Absoluta: La desviación media es la media de los valores absolutos de la desviación estándar.

$$D_{\bar{x}} = \frac{\sum_{i=1}^{N} |X_i - \bar{x}|}{N}$$

Diagrama de Caja: es un gráfico utilizado para representar una variable cuantitativa (variable numérica). El gráfico es una herramienta que permite visualizar, a través de los cuartiles, cómo es la distribución, su grado de asimetría, los valores extremos, la posición de la mediana, etc. Se compone de:

- Un rectángulo (caja) delimitado por el primer y tercer cuartil (Q1 y Q3). Dentro de la caja una línea indica dónde se encuentra la mediana (segundo cuartil Q2).
- Dos brazos, uno que empieza en el primer cuartil y acaba en el mínimo, y otro que empieza en el tercer cuartil y acaba en el máximo.
- Los datos atípicos (o valores extremos) que son los valores distintos que no cumplen ciertos requisitos de heterogeneidad de los datos.

Curva de Lorenz: es una forma gráfica de mostrar la distribución de la renta en una población. Por medio de ella se relacionan los porcentajes acumulados de población con porcentajes acumulados de la renta que esta población recibe.



El **índice Gini**, es un índice de concentración de la riqueza y equivale al doble del área de concentración. Su valor estará entre cero y uno. Cuanto más próximo a uno sea el índice Gini, mayor será la concentración de la riqueza; cuanto más próximo a cero, más equitativa es la distribución de la renta en ese país.

3. ¿Qué es PositTM y qué relación tiene con R Studio?

Posit es una modificación de marca o por ende el nuevo nombre que recibe la plataforma anterior que es conocida como RStudios, su misión principal es crear software de código abierto. Habrá un cambio de herramientas y productos comerciales: RStudio Connect = Posit Connect, Banco de trabajo RStudio = Banco de trabajo Posit, Administrador de paquetes de RStudio = Administrador de paquetes de Posit. En general posit y RStudio son lo mismo.

Bibliografía

1.3.-Medidas de Tendencia Central, de Dispersión, Simetría y Kurtosis. (s. f.). Recuperado 13 de agosto de 2022, de http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro19/13medidas _de_tendencia_central_de _dispersin_simetra_y _kurtosis.html

Cuantiles. (s. f.). Recuperado 13 de agosto de 2022, de http://www.uniquin.org/estadistica/cuantiles.html

Hidalgo, U. A. del E. de. (s. f.). Universidad Autónoma del Estado de Hidalgo: UAEH. Universidad Autónoma del Estado de Hidalgo. Recuperado 13 de agosto de 2022, de https://www.uaeh.edu.mx/

La distribución de la renta, la curva de Lorenz y el índice de Gini. (s. f.). Recupe-

rado 13 de agosto de 2022, de https://www.juntadeandalucia.es/averroes/centros-tic/14002996/helvia/aula/archivos/repositorio/250/271/html/economia/7/Lorenz-Gini.htm

M. Kelmansky, D. (s. f.). ANÁLISIS DE DATOS. Departamento de Matemática — Exactas — UBA. http://www.dm.uba.ar/materias/ analisis_de_datos/2008/1/ teoricas/Teor5.pdf

Medidas de dispersión. (s. f.). Recuperado 13 de agosto de 2022, de https: //www.universoformulas.com/ estadistica/descriptiva/ medidas-dispersion/

MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSION. (s. f.). Recuperado 13 de agosto de 2022, de https://eduteka.icesi.edu.co/proyectos.php/1/3053