UNIVERSIDAD
NACIONAL
DE COLOMBIA

Taller 1
Programación en lenguajes estadísticos

Estudiantes:
Cristian Javier García Pérez
Andrea Carolina Ropero Lozano
Deiver Adolfo Rodriguez Santiago
Neimer Jose Cervantes Lascarro

17/06/2022

1) Traducción de la sección “Elements of structured data” (págs. 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”.

Análisis exploratorio de datos.
Este capítulo se centra en el primer paso de cualquier proyecto de ciencia de datos: la exploración de los datos.La estadística clásica se centraba casi exclusivamente en la inferencia, un conjunto a veces complejo de procedimientos para sacar conclusiones sobre grandes poblaciones a partir de pequeñas muestras. En 1962, John W. Tukey (Figura 1-1) pidió una reforma de la estadística en su artículo seminal .[El] futuro del análisis de datos"[Tukey-1962]. Propuso una nueva disciplina científica llamada análisis de datos, que incluía la inferencia estadística como uno de sus componentes. Tukey forjó vínculos con las comunidades de ingeniería y ciencias de la computación (acuñó los términos bit, abreviatura de dígito binario, y software), y sus principios originales son sorprendentemente duraderos y forman parte de la base de la ciencia de los datos. El campo del análisis exploratorio de datos se estableció con el libro ya clásico de Tukey de 1977, Explore datos exploratorios [Tukey-1977]. Tukey presentó gráficos simples (por ejemplo, boxplots, scatde caja, gráficos de dispersión) que, junto con los estadísticos de resumen (media, mediana, cuartiles, etc.), ayudan a pintar una imagen de un conjunto de datos.Gracias a la disponibilidad de potencia informática y de programas de análisis de datos expresivos, el análisis exploratorio de datos ha evolucionado mucho, el análisis exploratorio de datos ha evolucionado mucho más allá de su alcance original. Los principales impulsores de esta disciplina han sido el rápido desarrollo de nuevas tecnologías, el acceso a más y el acceso a más datos, y el mayor uso del análisis cuantitativo en una variedad de disciplinas. David Donoho, profesor de estadística en la Universidad de Stanford y antiguo alumno de Tukey, explica que el análisis cuantitativo es una de las principales características de esta disciplina de Tukey, es autor de un excelente artículo basado en su presentación en el taller del centenario de Tukey en Princeton. Taller del centenario de Tukey en Princeton, Nueva Jersey [Donoho-2015]. Donoho remonta la génesis de la ciencia de datos hasta el trabajo pionero de Tukey en el análisis de datos.
Elementos de los datos estructurados.
Los datos provienen de muchas fuentes: mediciones de sensores, eventos, texto, imágenes y videos. 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, cada uno de los cuales contiene colores RGB (rojo, verde, azul). Los textos son secuencias de palabras y caracteres no verbales, a menudo organizados por secciones, subsecciones, etc. Los flujos de clics son secuencias de acciones de un usuario que interactúa con una aplicación o una página web. De hecho, uno de los propósitos principales de la ciencia de los datos es aprovechar este torrente de datos en bruto para convertirlos en información procesable. Para aplicar los conceptos estadísticos tratados 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 de una base de datos relacional o recogida para un estudio.
Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Los datos numéricos se presentan de 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 sólo toman un conjunto fijo de valores, como un tipo de valores, como un tipo de pantalla de televisión (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 sólo 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 las categorías están ordenadas; un ejemplo de una calificación numérica (1, 2, 3, 4 o 5).
¿Por qué nos molestamos en hacer una taxonomía de los tipos de datos? Resulta que, a efectos de análisis de datos y modelos predictivos, el tipo de datos es importante para ayudar a determinar el tipo de presentación visual, el análisis de datos o el modelo estadístico. De hecho, los software de ciencia de datos, como R y Python, utilizan estos tipos de datos para mejorar el rendimiento computacional. computacional. Más importante aún, el tipo de datos de una variable determina cómo el software manejará los cálculos para esa variable.

2) Definiciones de “Medidas de tendencia central y dispersión”:

Medidas de tendencia central (media aritmética, mediana y cuartiles, gráficos cuantil-cuantil, moda, media geométrica y media armónica).
A) Las Medidas de tendencia central: son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda.
B) Las medidas de Dispersión miden el grado de dispersión de los valores de la variable.
C) Media aritmética: Es la medida de posición más frecuentemente usada. Para calcular la media aritmética o promedio de un conjunto de observaciones se suman todos los valores y se divide por el número total de observaciones.

$$\mu = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Figura 1

D) Mediana: La mediana de una muestra de n observaciones la calculamos de la siguiente forma: ordenamos los datos de menor a mayor. Si el número de datos es impar, la mediana es el dato que ocupa la posición central. Si el número de datos es par, la mediana es el promedio de los dos datos centrales.
E) Cuartiles: El cuartil es cada uno de los tres valores que pueden dividir un grupo de números, ordenados de menor a mayor, en cuatro partes iguales.
F) Moda: La moda es el dato que ocurre con mayor frecuencia en el conjunto.
G) Media geométrica:En matemáticas y estadística, la media geométrica de una cantidad arbitraria de números es la raíz n-ésima del producto de todos los números; es recomendada para datos de progresión geométrica, para promediar razones, interés compuesto y números índice.
H) Media armónica: La media armónica de una cantidad finita de números es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos valores y es recomendada para promediar velocidades.
Medidas de dispersión (rango y rango intercuartil, desviación absoluta, varianza y desviación estándar, y coeficiente de variación).
Diagramas de caja.
El diagrama de caja es un gráfico propuesto por Tukey, para presentar los datos numéricos, especialmente útil para comparar distribuciones de varios conjuntos de observaciones y estaba basado en medidas robustas de posición y dispersión.
Medidas de concentración (curva de Lorenz y coeficiente Gini).
La curva de Lorenz y el coeficiente de Gini, son herramientas que se utilizan en el campo de la economía para medir la desigualdad de los ingresos de una población o sociedad.
En definitiva, tanto el Índice de Gini cómo la Curva de Lorenz, son métodos para identificar las desigualdades de ingresos en un determinado país o población. Por tanto, hemos de mencionar que cuánto más desarrollado está un país, más equidad suele haber en él. Es decir, el índice de Gini se acerca más a 0 y la Curva de Lorenz se acerca más a línea de perfecta igualdad.

3) ¿Qué es Posit™ y qué relación tiene con R Studio?

Posit™ es el nuevo nombre de la empresa Rstudio, su significado es “Postular” que es dar ideas u opiniones, con este nombre querían reflejar el trabajo de su agrupación o comunidad. La relación que tiene R studio con Posit™ es que es la misma empresa pero ahora con nuevo nombre que se le ha asignado a este entorno de desarrollo integrado (IDE), este lenguaje de programación R está dedicado a la computación estadística y gráficos, este lenguaje es muy útil ya que ayuda a muchas personas con el planteamiento de diferentes preguntas debido a los datos, esta entidad desarrolló un código abierto para que la ciencia de datos “code-first” (primeros datos) sea accesible para todas las personas con una Perspectiva elemental para el análisis y la comunicación.