Programación en lenguajes estadisticos

Santiago Silva, Maria Fernanda Castillo, Natalia Ballesteros

Universidad Nacional de Colmbia, sede de La Paz

11 de agosto de 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".

Figura 1-1. John Tukey, el eminente estadístico cuyas ideas desarrolladas hace más de 50 años forman la base de la ciencia de los datos

Elementos de los datos estructurados Los datos proceden de muchas fuentes: mediciones de sensores, eventos, textos, 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, 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 convertir este torrente de datos en bruto 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 recogidos 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 esto es 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. Más importante aún, el tipo de datos de una variable determina cómo el software de datos de una variable determina el modo en que el software gestionará los cálculos de esa variable.

Los ingenieros de software y los programadores de bases de datos pueden preguntarse por qué necesitamos los datos categóricos y ordinales para el análisis. Después de todo, las categorías no son más que una colección de valores de texto (o numéricos), y

Cuadro 1: Términos clave de los tipos de datos

Numéricos	Datos que se expresan en una escala numérica.
Continuo	Datos que pueden tomar cualquier valor en un intervalo.(Sinónimos: intervalo, float,numérico)
Discreto	Datos que sólo pueden tomar valores enteros, como los recuentos. (Sinónimos: entero,recuento)
Categórico	Datos que sólo pueden tomar un conjunto específico de valores que representan un conjunto de posibles categorías. (Sinónimos: enums, enumerado, factores, nominal)
Binario	Un caso especial de datos categóricos con sólo 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)

la base de datos subyacente se encarga automáticamente de 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 indique al software cómo de cómo deben comportarse los procedimientos estadísticos, como la elaboración de un gráfico o el ajuste de un modelo. En particular, los datos ordinales pueden representarse como un factor ordenado en R, preservando un ordenador en R, preservando un orden especificado por el usuario en gráficos, tablas y modelos. En Python, scikit-learn soporta datos ordinales con el sklearn.preprocessing.OrdinalEncoder.
- El almacenamiento y la indexación pueden ser optimizados (como en una base de datos relacional).
- Los posibles valores que puede tomar una variable categórica se imponen en el software (como un enum).

El tercer "beneficio"puede llevar a un comportamiento no deseado o inesperado: el comportamiento por defecto de las funciones de importación de datos en R (como el de la base de datos).columna de texto en un factor. Las operaciones posteriores sobre esa columna asumirán que los únicos valores permitidos para esa columna son los importados originalmente, y asignar un nuevo valor de texto introducirá una advertencia y producirá un NA (perdido valor). El paquete pandas en Python no realizará dicha conversión automáticamente. Sin embargo, puede especificar una columna como categórica explícitamente en la función read$_csv$.

Ideas claves

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

Otras lecturas

- La documentación de pandas describe los diferentes tipos de datos y cómo pueden ser manipulado en Python.
- Los tipos de datos pueden ser confusos, ya que los tipos pueden superponerse y la taxonomía en uno el software puede diferir del de otro. El sitio web R Tutorial cubre los taxonomía para 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; ver la guía de W3Schools para SQL.

Datos rectangulares es el término general para una matriz bidimensional con filas que indican ing registros (casos) y columnas que indican características (variables); marco de datos es el formato específico en R y Python. Los datos no siempre comienzan de esta forma: no estructurados. Los

datos (por ejemplo, texto) deben procesarse y manipularse para que puedan representarse como un conjunto de características en los datos rectangulares (ver "Elementos de 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 los casos,tareas de análisis y modelado de datos. **Datos rectangulares El marco de referencia típico para un análisis en ciencia de datos es un marco de datos rectangular, objeto, 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 ing registros (casos) y columnas que indican características (variables); marco de datos es el formato específico en R y Python. Los datos no siempre comienzan de esta forma: no estructurados. Los datos (por ejemplo, texto) deben procesarse y manipularse para que puedan representarse como un conjunto de características en los datos rectangulares (ver "Elementos de 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 los casos,tareas de análisis y modelado de datos.

2. Definiciones de "Medidas de tendencia central y dispersion"

2.1. Medidas de tendencia central (media arıtmetica, mediana y cuantiles, graficos cuantil-cuantil, moda, media geometrica y media harmonica).

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. Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable. Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y su dispersión. Los procedimientos para obtener las medidas estadísticas difieren levemente dependiendo de la forma en que se encuentren los datos. Si los datos se encuentran ordenados en una tabla estadística diremos que se encuentran "agrupados" y si los datos no están en una tabla hablaremos de datos "no agrupados". Según este criterio, haremos primero el estudio de las medidas estadísticas para datos no agrupados y luego para datos agrupados.

Medidas estadísticas en datos no agrupados

Medidas de tendencia central

Promedio o media aritmética

La medida de tendencia central más conocida y utilizada es la media aritmética o promedio aritmético. Se representa por la letra griega μ cuando se trata del promedio del universo o población y por $\bar{Y}$ (léase Y barra) cuando se trata del promedio de la muestra. Es importante destacar que μ es una cantidad fija mientras que el promedio

de la muestra es variable puesto que diferentes muestras extraídas de la misma población tienden a tener diferentes medias. La media se expresa en la misma unidad que los datos originales: centímetros, horas, gramos, etc.

Mediana

Otra medida de tendencia central es la mediana. La mediana es el valor de la variable que ocupa la posición central, cuando los datos se disponen en orden de magnitud. Es decir, el $50\,\%$ de las observaciones tiene valores iguales o inferiores a la mediana y el otro $50\,\%$ tiene valores iguales o superiores a la mediana. Si el número de observaciones es par, la mediana corresponde al promedio de los dos valores centrales. Por ejemplo, en la muestra $3, 9, 11, 15$, la mediana es $(9 + 11)/2 = 10$.

Moda

La moda de una distribución se define como el valor de la variable que más se repite. En un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto del gráfico. Una muestra puede tener más de una moda.

Cuartiles

Los cuartiles son valores que dividen una muestra de datos en cuatro partes iguales. Utilizando cuartiles se puede evaluar rápidamente la dispersión y la tendencia central de un conjunto de datos, que son los pasos iniciales importantes para comprender sus datos.

Gráficos cuantil-cuantil

Un gráfico Cuantil-Cuantil permite observar cuán cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos.

Moda

La moda de una distribución se define como el valor de la variable que más se repite. En un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto del gráfico. Una muestra puede tener más de una moda.

Media geométrica

La media geométrica (MG) de un conjunto de números estrictamente positivos (X1, X2,. . . ,XN) es la raíz N-ésima del producto de los N elementos.

$$MG = \sqrt[n]{x_1, x_2 ... x_n}$$

Todos los elementos del conjunto tienen que ser mayores que cero. Si algún elemento fuese cero (Xi=0), entonces la MG sería 0 aunque todos los demás valores estuviesen alejados del cero.

media armónica

La media armónica (H) de un conjunto de elementos no nulos (X1, X2,. . . ,XN) es el recíproco de la suma de los recíprocos (donde 1/Xi es el recíproco de Xi)) multiplicado por el número de elementos del conjunto (N).

$$H = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_N}}$$

2.2. Medidas de dispersión (rango y rango intercuartil, desvizaón absoluta, varianza y desviación est ándar, y coeficiente de varacion).

Las medidas de dispersión entregan información sobre la variación de la variable. Pretenden resumir en un solo valor la dispersión que tiene un conjunto de datos. Las medidas de dispersión más utilizadas son: Rango de variación, Varianza, Desviación estándar, Coeficiente de variación.

Rango

El rango es un valor numérico que sirve para manifestar la diferencia entre el valor máximo y el valor mínimo de una muestra poblacional. A través del rango se puede observar la dispersión total en una muestra en concreto. Este parámetro estadístico es especialmente utilizado en finanzas, ya que resulta de gran utilidad para observar el tamaño que podría adquirir una variación.

Rango intercuartil

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).

desviación absoluta

Esta desviación muestra la variación que tiene cada uno de los datos de un grupo con respecto a su media aritmética, lo que nos permitirá determinar que tan homogéneo es el grupo de datos.

Varianza

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones.

Rango de variación

Se define como la diferencia entre el mayor valor de la variable y el menor valor de la variable. La mejor medida de dispersión, y la más generalizada es la varianza,

o su raíz cuadrada, la desviación estándar. La varianza se representa con el símbolo σ^2 (sigma cuadrado) para el universo o población y con el símbolo s^2 (s cuadrado), cuando se trata de la muestra. La desviación estándar, que es la raíz cuadrada de la varianza, se representa por Σ (sigma) cuando pertenece al universo o población y por "s", cuando pertenece a la muestra. σ^2 y σ son parámetros, constantes para una población particular; s2 y s son estadígrafos, valores que cambian de muestra en muestra dentro de una misma población. La varianza se expresa en unidades de variable al cuadrado y la desviación estándar simplemente en unidades de variable.

Desviación estándar

La desviación estándar es una medida de la dispersión de los datos, cuanto mayor sea la dispersión mayor es la desviación estándar, si no hubiera ninguna variación en los datos, es decir, si fueran todos iguales, la desviación estándar sería cero. La desviación estándar cuantifica la dispersión alrededor de la media aritmética. Informa de la media de distancias que tienen los datos respecto de su media aritmética.

Coeficiente de variación

Es una medida de la dispersión relativa de los datos. Se define como la desviación estándar de la muestra expresada como porcentaje de la media muestral. Es de particular utilidad para comparar la dispersión entre variables con distintas unidades de medida. Esto porque el coeficiente de variación, a diferencia de la desviación estándar, es independiente de la unidad de medida de la variable de estudio.

2.3. Diagramas de caja.

El 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.

2.4. 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.

Las medidas de concentración nos informan de la concentración de la distribución , entendida en un sentido distinto al de la antinomia "dispersión/ concentración": miden lo que podríamos llamar la concentración en sentido .económico":miden el mayor o menor "grado de igualdad en el reparto de la totalidad de los valores de la variable. De esta manera si una pequeña parte de la población (unos pocos individuos) tiene una gran parte del total de la variable (renta, salario, capital, etc.), la variable estará muy concentrada (en pocas manos).Sin embargo, si se guardan las proporciones entre individuos y parte del total que se reparten la distribución será igualitaria, homogénea, poco o nada concentrada

3. ¿Que es PositTM y que relacion tiene con R Studio?

RStudio tiene un nuevo nombre: Posit. Este es un gran cambio, ¿por qué Posit? Posit es una palabra real (se abre en una nueva pestaña) que significa proponer una idea para la discusión. Los científicos de datos pasan gran parte de su día planteando afirmaciones que luego evalúan con datos. Al considerar un nuevo nombre para la empresa, querían algo que reflejara tanto el trabajo que realiza nuestra comunidad (como la aspiración científica de construir niveles de conocimiento y comprensión cada vez mayores.

Están en el inicio de una nueva fase de desarrollo de RStudio. Para la primera fase, tomaron la decisión potencialmente confusa de nombrar a nuestra compañía con el nombre de nuestro IDE que inicialmente estaba enfocado a los usuarios de R. Mantuvimos ese nombre incluso cuando la oferta creció a mucho más que un IDE, y sirvió a muchos lenguajes aparte de R. Si bien eso tenía sentido en ese momento, se ha vuelto cada vez más difícil mantener ese nombre a medida que nuestra carta se ha ampliado: identificar el mejor IDE para la ciencia de datos con R.

¿Qué significa el nuevo nombre para el software comercial? En muchos sentidos, nada: nuestros productos comerciales han soportado Python durante más de 2 años. Pero vamos a cambiar el nombre a Posit Connect, Posit Workbench, y Posit Package Mana-

ger para que sea más fácil para la gente entender que apoyamos más que sólo R. ¿Qué pasa con nuestro software de código abierto? Del mismo modo, no está cambiando mucho: nuestro software de código abierto es y seguirá siendo predominantemente para R. Dicho esto, en los últimos años ya hemos estado invirtiendo en otros lenguajes como reticulate (llamando a Python desde R), características de Python para el IDE, y soporte para Python y Julia dentro de Quarto. Puedes esperar ver más experimentos multilingües en el futuro.