TALLER 1

Programación en Lenguajes Estadisticos

Kevin José Cáceres Contreras
Wilfran Ballesteros Diaz

24 de agosto de 2022

John Tukey, el eminente estadístico cuyas ideas se desarrollaron hace más de 50 añosformar la base de la ciencia 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 esto los 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 sin palabras acters, a menudo organizados por secciones, subsecciones, etc. Los clickstreams son sequences de acciones de un usuario que interactúa con una aplicación o una página web. De hecho, un importante El desafío de la ciencia de datos es aprovechar este torrente de datos en bruto en información procesable. mation. Para aplicar los conceptos estadísticos cubiertos en este libro, datos brutos no estructurados debe ser procesado y manipulado en una forma estructurada. Uno de los más comunes Formularios de datos estructurados es una tabla con filas y columnas, ya que pueden surgir datos de una base de datos relacional o ser recolectado para un estudio.

Hay dos tipos básicos de datos estructurados: numéricos y categóricos. Datos numéricos viene en dos formas: *continua*, como la velocidad del viento o la duración del tiempo, y *discreta*, como el recuento de la ocurrencia de un evento. Los datos *categóricos* solo toman un conjunto fijo de valores, como un tipo de pantalla de TV (plasma, LCD, LED, etc.) o un nombre de estado (Alabama, Alaska, etc.). Los datos *binarios* son un caso especial importante de datos categóricos que toma solo uno de dos valores, como 0/1, sí/no o verdadero/falso. Otro tipo útil de datos categóricos son datos *ordinales* en los que se ordenan las categorías; un ejemplo de esta es una calificación numérica (1, 2, 3, 4 o 5).

¿Por qué nos molestamos con una taxonomía de tipos de datos? Resulta que

para los propósitos del análisis de datos y el modelado predictivo, el El tercer "beneficio"puede conducir a un comportamiento no deseado o inespe- rado: el valor predeterminado el comportamiento de las funciones de importa- ci on de datos en R (por ejemplo, read.csv) es convertir autom aticamente una columna de texto en un factor. Las operaciones subsiguientes en esa columna su- pondr an que los unicos valores permitidos para esa columna son los importados originalmente, y asignar un nuevo valor de texto introducir a una advertencia y producir a un NA (falta valor). El paquete pandas en Python no realizar a dicha conversi on autom aticamente. Sin embargo, puede especificar una columna como categ orica expl ıcitamente en la funci on readcsv. Ideas Claves Los datos generalmente se clasifican en software por tipo. Los tipos de datos incluyen num ericos (continuos, discretos) y categ oricos (binarios, ordinal). Lecturas Adicionales La documentaci on de pandas describe los diferentes tipos de datos y c omo pueden ser manipulado en Python. Los tipos de datos pueden ser confusos, ya que los tipos pueden super- ponerse y la taxonom ıa en uno el software puede diferir del de otro. El sitio web de R Tutorial cubre el taxonom ıa para R. La documentaci on de pandas describe los diferentes tipos de datos y c omo se pueden manipular en Python. Las bases de datos son m as detalladas en su clasificaci on de tipos de datos, incorporando consideraciones de niveles de precisi on, campos de longitud fija o variable, y m as; ver la gu ıa de W3Schools para SQL. Datos Rectangulares El marco de referencia t ıpico para un an alisis en ciencia de datos es un dato rectangular objeto, como una hoja de c alculo o una tabla de base de datos. Datos rectangulares es el t ermino general para una matriz bidimensional con filas que indican registros (casos) y columnas que indican caracter ısticas (variables); el marco de datos es el formato espec ífico en R y Python. Los da- tos no siempre comienzan de esta forma: no estructurados datos (por ejemplo, texto) debe procesarse y manipularse para que pueda representarse como un conjunto de caracter ısticas en los datos rectangulares (ver " Elementos de Da- tos Estructurados" en la p agina 2).Los datos en las bases de datos relacionales deben extraerse y colocarse en una sola tabla para la mayor ıa tareas de an alisis y modelado de datos. 3 Medidas de tendencia central Las medidas de tendencia central son valores que se ubican al centro de un conjunto de datos ordenados seg un su magnitud. Generalmente se utilizan 4 de estos valores tambi en conocidos como estadigrafos, la media aritm etica, la mediana, la moda y al rango medio. 1. Media aritmetica: 2. Mediana: Otra medida de tendencia central es la mediana. La mediana es el valor de la variable que ocupa la posici on central, cuando los datos se disponen en orden de magnitud. Es decir, el 50 tiene valores iguales o inferiores a la mediana y el otro 50 iguales o superiores a la mediana. Si el n umero 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. 3. Cuantiles: 4. Moda: La moda de una distribuci on se define como el valor de la variable que m as se repite. En un pol ıgono de frecuencia la moda corresponde al valor de la variable que est a bajo el punto m as alto del gr afico. Una muestra puede tener m as de una moda. 5. La media geom etrica: La media geom etrica se calcula como un pro- ducto conjunto. Es decir, que todos los valores se multiplican entre s

ı. De modo que si uno de ellos fuera cero, el producto total ser ıa cero. Por ello, debemos siempre tener en cuenta que a la hora de calcular la media geom etrica necesitamos n umeros que sean unicamente positivos. Uno de sus principales usos es para calcular medias sobre porcentajes, pues su c alculo ofrece unos resultados m as adaptados a la realidad. La f ormula de la media geom etrica es la siguiente: N x1 · x23 · ... · xN 6. La media arm onica: La media arm onica es igual al n umero de elementos de un grupo de cifras entre la suma de los inversos de cada una de estas cifras. La f ormula de la media arm onica (H) de un conjunto de n umeros x1, x2, x3, . . . , xn, es la siguiente: H = N 1 x1 + 1 x2 +...+ 1 xn RangoElrangoesunvalornum ericoqueindicaladif erenciaentreelvalorm aximoyelm ınimodeunapoblaci on 4 tipo de datos es importante para ayudar a determinar el tipo de visualización visual, análisis de datos o modelo estadístico. De hecho,la ciencia de datos El software, como R y *Python*, utiliza estos tipos de datos para mejorar la performance. Más importante aún, el tipo de datos para una variable determina cómo el software controlar cálculos para esa variable.

Términos clave para tipos de datos

1. **Númerico**
   Datos que se expresan en una escala numérica.
   - **Continuo:** Datos que pueden tomar cualquier valor en un intervalo.
   - **DiscretoDatos:** que solo pueden tomar valores enteros, como recuentos. (*Sinónimos*: entero,contar)
2. **Categórico**
   Datos que pueden tomar solo un conjunto específico de valores que representan un conjunto de posibles categorías. (*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 análisis. Después de todo, las categorías son meramente un 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 qué tan estadístico 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 conducir a un comportamiento no deseado o inesperado: el valor predeterminado el comportamiento 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 importados originalmente, y asignar un nuevo valor de texto introducirá una advertencia y producirá un NA (falta 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 generalmente se clasifican en software por tipo.
- Los tipos de datos incluyen numéricos (continuos, discretos) y categóricos (binarios, ordinal).

Lecturas Adicionales

- 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 de R Tutorial cubre el 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

El marco de referencia típico para un análisis en ciencia de datos es un *dato 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 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: no estructurados datos (por ejemplo, texto) debe procesarse y manipularse para que pueda representarse como un conjunto de características en los datos rectangulares (ver " Elementos de Datos Estructurados" en la página 2).Los datos en las bases de datos relacionales deben extraerse y colocarse en una sola tabla para la mayoría tareas de análisis y modelado de datos.

1. Medidas de tendencia central

Las medidas de tendencia central son valores que se ubican al centro de un conjunto de datos ordenados según su magnitud. Generalmente se utilizan 4 de estos valores también conocido como estadigrafos, la media aritmética, la mediana, la moda y al rango medio.

- **Media aritmetica:**La media aritmética es el valor promedio de las muestras y es independiente de las amplitudes de los intervalos. Se simboliza como y se encuentra sólo para variables cuantitativas. Se encuentra sumando todos los valores y dividiendo por el número total de datos.
  La fórmula general para N elementos es:

$$x = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

- **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.
- **Cuantiles:** Los cuartiles son medidas estadísticas de posición que tienen la propiedad de dividir la serie estadística en cuatro grupos de números iguales de términos.

- **gráficos cuantil-cuantil:**

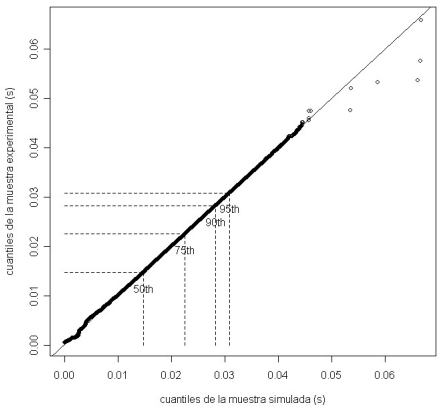

Figura 1: cuantil-cuantil

- **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.
- **La media geométrica:** La media geométrica se calcula como un producto conjunto. Es decir, que todos los valores se multiplican entre sí. De modo que si uno de ellos fuera cero, el producto total sería cero. Por ello, debemos siempre tener en cuenta que a la hora de calcular la media geométrica necesitamos números que sean únicamente positivos.

  Uno de sus principales usos es para calcular medias sobre porcentajes, pues su cálculo ofrece unos resultados más adaptados a la realidad. La fórmula de la media geométrica es la siguiente:

$$\sqrt[N]{x_1 \cdot x_{23} \cdot ... \cdot x_N}$$

- **La media armónica:** La media armónica es igual al número de elementos de un grupo de cifras entre la suma de los inversos de cada una de estas cifras. La fórmula de la media armónica (H) de un conjunto de números $x_1, x_2, x_3, \ldots, x_n$, es la siguiente:

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

  cabe destacar que N es el número de elementos sobre los cuales se calcula la media.

2. Medidas de dispersión:

- **Rango:** El rango es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de una población o muestra estadística. Su fórmula es:
  $R = Máx_x – Mín_x$

  Donde: R → Es el rango.
  Máx → Es el valor máximo de la muestra o población.
  Mín → Es el valor mínimo de la muestra o población estadística.
  x → Es la variable sobre la que se pretende calcular esta medida

- **Rango intercuartil:** el rango intercuartílico es la diferencia entre el penúltimo y el primer cuartil de una distribución utilizado en el diagrama de caja. Generalmente utilizado en el diagrama de caja que utiliza la mediana como medida central.
  La forma abreviada de nombrar al rango intercuartílico es RIC o RQ. El rango intercuartil utiliza la mediana como medida central. Entonces, el resultado del rango intercuartil será próximo a la mediana o segundo cuartil (Q2) si hay pocos valores extremos.
  Fórmula del rango intercuartílico:
  Sabiendo que el rango intercuartil es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), entonces, simplemente tenemos que hacer la diferencia entre ambos valores.
  RIC = Q3 – Q1

- **desviación absoluta:** como recorrido, desviación media, varianza y desviación típica, que se usanen los análisis estadísticos generales.
  Medidas de dispersión relativa: que determinan la dispersión de la distribución estadística independientemente de las unidades en que se exprese la variable.

- **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. Su fórmula es la siguiente:

$$\sigma^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{N}$$

  X → Variable sobre la que se pretenden calcular la varianza.
  xi → Observación número i de la variable X. i puede tomará valores entre 1 y n.
  N → Número de observaciones.
  x → Es la media de la variable X.

- **Desviación estándar ó Desviación típica:** La desviación típica es otra medida que ofrece información de la dispersión respecto a la media. Su cálculo es exactamente el mismo que la varianza, pero realizando la raíz cuadrada de su resultado. Es decir, la desviación típica es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{x})^2}{N}}$$

X → Variable sobre la que se pretenden calcular la varianza.
xi → Observación número i de la variable X. i puede tomará valores entre 1 y n.
N → Número de observaciones. x → Es la media de la variable X.

- **Coeficiente de variación:** Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la media del conjunto y por lo general se expresa en porcentaje para su mejor comprensión.

X → Variable sobre la que se pretenden calcular la varianza.
x → Desviación típica de la variable X.
— x — → Es la media de la variable X en valor absoluto con x 0

3. ## Diagramas de caja:

Los diagramas de caja le permiten visualizar y comparar la distribución y la tendencia central de valores numéricos mediante sus cuartiles. Los cuartiles representan un método para dividir valores numéricos en cuatro grupos iguales basados en cinco valores clave: mínimo, primer cuartil, mediana, tercer cuartil y máximo.

La parte de la caja del diagrama siguiente ilustra el 50 por ciento medio de los valores de los datos, también conocido como rango intercuartílico o IQR. Le media de los valores se representa como la línea que divide la caja por la mitad. El IQR ilustra la variabilidad en un conjunto de valores. Un IQR grande indica una amplia dispersión de los valores, mientras que un IQR más pequeño indica que la mayoría de los valores quedan hacia el centro. Los diagramas de caja también ilustran los valores mínimos y máximos de los datos mediante bigotes, o líneas, que se extienden desde la caja y, opcionalmente, valores atípicos como puntos que se extienden más allá de los bigotes.

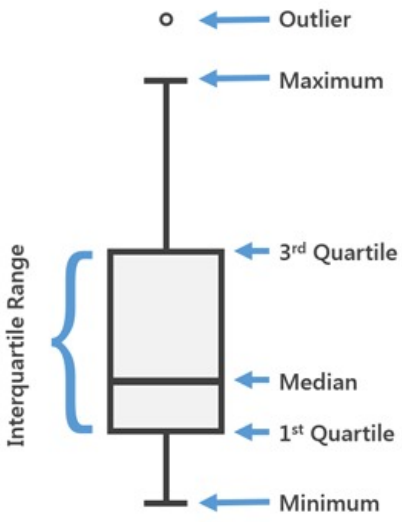

Figura 2: Diagrama de caja

4. Medidas de concentración:

- **Curva de Lorenz:** es una forma gráfica de mostrar la distribución de la renta en una población. En ella se relacionan los porcentajes acumulados de población con porcentajes acumulados de la renta que esta población recibe. En el eje de abcisas se representa la población .ºrdenada"de forma que los percentiles de renta más baja quedan a la izquierda y los de renta más alta quedan a la derecha. El eje de ordenadas representa las rentas.

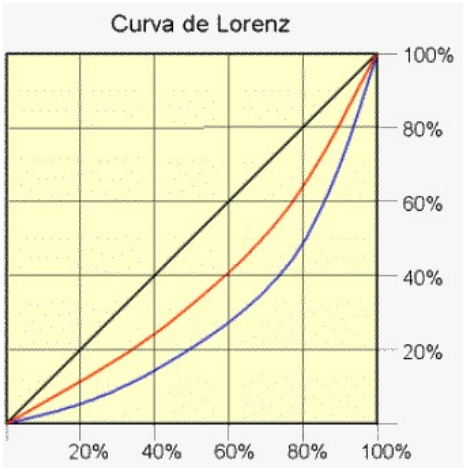

Figura 3: Curva de lorenz

En la gráfica se muestran como ejemplo la representación de dos países imaginarios, uno en azul y otro en rojo. La distribución de la renta en el país azul es más desigual que en el país rojo. En el caso del país azul, el cuarenta por ciento más pobre de la población recibe una renta inferior al veinte por ciento del total del país. En cambio, en el país rojo, el cuarenta por ciento más pobre recibe más del veinte por ciento de la renta. La línea diagonal negra muestra la situación de un país en el que todos y cada uno de los individuos obtuviese exactamente la misma renta; sería la equidad absoluta. Cuanto más próxima esté la curva de Lorenz de la diagonal, más equitativa será la distribución de la renta de ese país.

- **Coeficiente Gini:** 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.

¿Qué es Posit^{TM} y qué relación tiene con R Studio?

Posit, PBC es el nuevo nombre corporativo de la empresa anteriormente conocida como RStudio, PBC. Es un cambio de marca que refleja la expansión a Python y VS Code, entre otras cosas.

El nuevo nombre abre la puerta a la empresa para salir de su encasillamiento superficial como una empresa R-only. Posit puede continuar haciendo crecer el IDE de RStudio mientras se aleja de su amada sombra que consolidó a 'RStudio'como el principal IDE de R.

Con el cambio de marca, Posit se está moviendo más rápido hacia lo que RStudio ya estaba en camino de convertirse:

El puente hacia una mejor ciencia de datos.

Con un nuevo nombre y tentadores planes para el futuro, ese puente está abierto para expansión y aquí en Appsilon , estamos muy emocionados.

Los servicios y el software comercial de Posit siguen siendo las mismas herramientas poderosas que antes; con el mismo soporte para más que solo R. Si su empresa está buscando escalar y construir mejores productos de datos, Posit es la respuesta.

Durante los últimos años, Posit (anteriormente RStudio) ha estado cambiando de herramientas exclusivas de R a un ecosistema agnóstico del lenguaje. Para nuestro disfrute, hemos visto que RStudio IDE crece para ser más compatible con Python y que el ecosistema de ciencia de datos de Posit se convierte en Ün hogar único para R Python ".

El nombre que era sinónimo de desarrollo R de código abierto está cambiando de marca para representar mejor al negocio en su conjunto.

Entonces, ¿qué significa esto para los usuarios de productos RStudio? Bueno, además de un futuro más brillante lleno de más capacidades, no mucho a corto plazo. Habrá un cambio de marca 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 Pero en general, Posit no se alejará de R. Así que no se preocupe, RStudio IDE seguirá existiendo y los líderes en el desarrollo de R de código abierto no se están desacelerando.

En un comunicado de prensa de Sharon Machlis , se citó a Hadley Wickham, científico jefe de RStudio, diciendo: "No vamos a pasar de R a Python".

Hadley nos tranquilizó a todos al continuar: "... No voy a dejar de escribir código Rz luego declaró: "No voy a aprender Python".

Y aunque RStudio busca equilibrar la proporción de ingenieros que trabajan en R con otros desarrollos a lo largo del tiempo, la mayoría del trabajo seguirá estando relacionado con R.