Programacion en Lenguajes Estadistico

Euller Francisco Alvarado Vargas ealvarado@unal.edu.co
Albeiro Junior Diaz Oquendo adiazoq@unal.edu.co
Ruben Dario Ariza rariza@unal.edu.co

15,Agosto, 2022

1. **Traduccion:**

Estadística práctica para Científicos de datos

Los métodos estadísticos son una parte clave de la ciencia de datos, sin embargo, pocos científicos de datos tienen capacitación estadística formal. Los cursos y libros sobre estadísticas básicas rara vez cubren el tema desde la perspectiva de la ciencia de datos. La segunda edición de esta popular guía agrega ejemplos completos en Python, brinda orientación práctica sobre la aplicación de métodos estadísticos a la ciencia de datos, le indica cómo evitar su mal uso y le brinda consejos sobre lo que es importante y lo que no. Muchos recursos de ciencia de datos incorporan métodos estadísticos pero carecen de una perspectiva estadística más profunda. Si está familiarizado con los lenguajes de programación R o Python y tiene cierta exposición a las estadísticas, esta referencia rápida cierra la brecha en un formato accesible y confiable.

Con este libro, aprenderá:

- Por qué el análisis exploratorio de datos es un paso preliminar clave en la ciencia de datos.
- Cómo el muestreo aleatorio puede reducir el sesgo y generar un conjunto de datos de mayor calidad, incluso con big data.
- Cómo los principios del diseño experimental dan respuestas definitivas a las preguntas.
- Cómo usar la regresión para estimar resultados y detectar anomalías.
- Técnicas de clasificación clave para predecir a qué categorías pertenece un registro.
- Métodos de aprendizaje automático estadístico que "aprenden" de los datos.
- Métodos de aprendizaje no supervisados para extraer significado de datos no etiquetados.

"Este libro no es otro libro de texto de estadística ni un manual de aprendizaje automático. Es mucho mejor: establece la conexión entre términos y principios estadísticos útiles y la jerga y las prácticas actuales de minería de datos, con explicaciones claras y muchos ejemplos. Esta es una excelente referencia para principiantes y veteranos de la ciencia de datos". - Galit Shmueli autor principal de la serie más vendida Data Mining for Business Analytics y profesor distinguido, Universidad Nacional Tsing Hua, Taiwán

Peter Bruce es el fundador del instituto de Educación Estadística en Statistics.com. Andrew Bruce es científico investigador principal en Amazon y tiene más de 30 años de experiencia en estadística y ciencia de datos. Peter Gedeck es científico de datos sénior en Collaborative Drug Discovery y desarrolla algoritmos de aprendizaje automático para predecir las propiedades de los fármacos candidatos.

SEGUNDA EDICIÓN

Estadística práctica para Científicos de datos Más de 50 conceptos esenciales con R y Python

Peter Bruce, Andrew Bruce y Peter Gedeck

 Impreso en los Estados Unidos de América. Publicado por O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. Los libros de O'Reilly pueden adquirirse para uso educativo, comercial o promocional. Las ediciones en línea están también están disponibles para la mayoría de los títulos (http://oreilly.com). Para más información, póngase en contacto con nuestro departamento de ventas departamento de ventas: 800-998-9938 o corporate@oreilly.com. Editor: Nicole Tache Editor de producción: Kristen Brown Correctora: Piper Editorial Corrector de pruebas: Arthur Johnson Indexador: Ellen Troutman-Zaig Diseñador de interiores: David Futato Diseñador de la cubierta: Karen Montgomery Ilustrador: Rebecca Demarest Mayo de 2017: Primera edición Mayo de 2020: Segunda edición Historial de revisiones de la segunda edición 2020-04-10: Primera edición Consulte http://oreilly.com/catalog/errata.csp?isbn=9781492072942 para conocer los detalles del lanzamiento. El logotipo de O'Reilly es una marca registrada de O'Reilly Media, Inc. Estadística práctica para científicos de datos, la imagen de portada y la imagen comercial relacionada son marcas comerciales de O'Reilly Media, Inc.

2. Definiciones

Medidas de tendencia central
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.

Media aritmética
La media aritmética es lo que se conoce como media al uso. Sumando todos los valores y dividiéndolos entre la cantidad de observaciones. Por ejemplo, imaginemos que queremos saber a cuantos trozos de pastel tocamos. Hay 10 trozos y somos 5 personas. Si lo repartimos a partes iguales, el resultado será de 2 trozos por persona. Sin darnos cuenta, acabamos de calcular una media aritmética.
https://economipedia.com/definiciones/media-aritmetica.html

La mediana y los cuartiles
La mediana y los cuartiles, como la media aritmética, sólo se pueden calcular cuando la variable es cuantitativa. La mediana, es el valor que ocupa la posición central una vez ordenados los datos en orden creciente, es decir, el valor que es mayor que el 50La mediana divide la distribución en dos partes con igual número de datos, si la dividimos en cuatro partes obtenemos los cuartiles, 1^{o}, 2^{o} y 3^{o}, que se indican respectivamente Q1, Q2 y Q3. Ordenados los datos, el primer cuartil, es mayor que el 25 de estos; el tercer cuartil, mayor que el 75, y el segundo coincide con la mediana.

Gráficos cuantil-cuantil
Los diagramas cuantil-cuantil son una herramienta de exploración utilizada para evaluar las similitudes entre la distribución de una variable numérica y una distribución normal, o entre las distribuciones de dos variables numéricas. Existen dos tipos de diagramas cuantil-cuantil: diagramas cuantil-cuantil normales y diagramas cuantil-cuantil generales. Los diagramas cuantil-cuantil normales se construyen trazando los cuantiles de una variable numérica respecto de los cuantiles de una distribución normal. Los diagramas cuantil-cuantil generales trazan los cuantiles de una variable numérica respecto de los cuantiles de una segunda variable numérica. Si las distribuciones de los cuantiles comparados son idénticas, los puntos del diagrama formarán una línea recta de 45 grados. Cuanto más lejos se desvíen los puntos del diagrama de una línea recta, menos similares serán las distribuciones comparadas.
https://pro.arcgis.com/es/pro-app/latest/help/analysis/geoprocessing/charts/qq-plot.htm

Moda
La moda estadística es aquel valor que, dentro de un conjunto de datos, se

repite el mayor número de veces. La determinación de la moda estadística en un conjunto de datos que no están agrupados no requiere ningún tipo de cálculo, sino tan solo el conteo de las variables. Otra forma de determinar la moda en datos no agrupados consiste en verificar cuál es el valor de mayor frecuencia en una tabla de frecuencias absolutas. La moda estadística es aplicable tanto para datos de información cualitativa como cuantitativa. Tipos de moda estadística La moda estadística se clasifica de la siguiente manera:
Moda unimodal: tipo de moda estadística en la cual un único valor se repite el mayor número de veces dentro de un conjunto de datos.
Moda bimodal: tipo de moda estadística en la que 2 valores diferentes presentan el mismo número máximo de repeticiones, dentro de un conjunto de datos.
Moda multimodal: tipo de moda estadística en la que 3 o más valores diferentes presentan el mismo número máximo de repeticiones dentro de un conjunto de datos.
https://enciclopediaeconomica.com/moda-estadistica/

Media geométrica

La media geométrica es un tipo de media que se calcula como la raíz del producto de un conjunto de números estrictamente positivos. 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. Fórmula de la media geométrica La fórmula de la media geométrica es la siguiente:

https://economipedia.com/definiciones/media-geometrica.html

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. En otras palabras, la media armónica es una medida estadística recíproca a la media aritmética, que es la suma de un conjunto de valores entre el número de observaciones. Fórmula de la media armónica La fórmula de la media armónica (H) de un conjunto de números x1, x2, x3, . . . , xn, es la siguiente:

Cabe destacar que N es el número de elementos sobre los cuales se calcula la media.
https://economipedia.com/definiciones/media-armonica.html

Medidas de dispersión

Las medidas de dispersión tratan, a través del cálculo de diferentes fórmulas, de arrojar un valor numérico que ofrezca información sobre el grado de variabilidad de una variable. En otras palabras, las medidas de dispersión son números que indican si una variable se mueve mucho, poco, más o menos que otra. La razón

de ser de este tipo de medidas es conocer de manera resumida una característica de la variable estudiada. En este sentido, deben acompañar a las medidas de tendencia central.
https://economipedia.com/definiciones/medidas-de-dispersion.html

Rango

El rango es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de una población o muestra estadística. El rango suele ser utilizado para obtener la dispersión total. Es decir, si tenemos una muestra con dos observaciones: 10 y 100 pesos, el rango será de 90 pesos.
https://economipedia.com/definiciones/rango-estadistica.html

Rango intercuartil

El RIQ describe el 50https://es.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6th/v/calculating-interquartile-range-iqr

Desviación absoluta

En estadística la desviación absoluta promedio o, sencillamente desviación media o promedio de un conjunto de datos es la media de las desviaciones absolutas y es un resumen de la dispersión estadística.1 Se expresa, de acuerdo con esta fórmula:

https://es.wikipedia.org/wiki/Desviaci

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. También se puede calcular como la desviación típica al cuadrado. Dicho sea de paso, entendemos como residuo a la diferencia entre el valor de una variable en un momento y el valor medio de toda la variable.
Fórmula de la varianza

https://economipedia.com/definiciones/varianza.html: :text=La

Desviación estándar

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos. El símbolo (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido. La desviación estándar se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.
https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistics/basic-

statistics/supporting-topics/data-concepts/what-is-the-standard-deviation/

Coeficiente de variación
El coeficiente de variación, también denominado coeficiente de variación de Pearson, es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos. Es decir, nos informa al igual que otras medidas de dispersión, de si una variable se mueve mucho, poco, más o menos que otra. Fórmula de coeficiente de variación

https://economipedia.com/definiciones/coeficiente-de-variacion.html

Diagramas de caja
Un diagrama de caja, del inglés, boxplot, es una representación de una variable cuantitativa o categórica con el propósito de identificar rápidamente los cuartiles del conjunto de datos. En otras palabras, un diagrama de caja es un gráfico que representa una variable cuantitativa o cualitativa a través de los cuartiles. En estadística, es una herramienta útil para representar conjuntos de datos tanto discretos como continuos. Es importante tener en cuenta que las variables cualitativas o que pretenden representar un orden o una categoría siempre tienen que ir ligadas a un índice numérico mayor que 0 para que puedan aparecer en el gráfico y se puedan calcular los estadísticos correspondientes.
https://economipedia.com/definiciones/diagrama-de-caja.html

Medidas de concentración
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.
https://www.uv.es/ceaces/base/descriptiva/concentrac.htm

Curva de Lorenz
La curva de Lorenz es una representación gráfica de la desigualdad en el reparto de la renta existente en un determinado territorio (normalmente un país). En ella, se sitúa en el eje X los acumulados de población (P) expresados en tanto por ciento y en el eje Y los acumulados de renta (Q) expresados en tanto por ciento.

Coeficiente Gini

El coeficiente de Gini es una medida de la desigualdad ideada por el estadístico italiano Corrado Gini. Normalmente se utiliza para medir la desigualdad en los ingresos, dentro de un país, pero puede utilizarse para medir cualquier forma de distribución desigual. El coeficiente de Gini es un número entre 0 y 1, donde 0 se corresponde con la perfecta igualdad (todos tienen los mismos ingresos) y donde el valor 1 se corresponde con la perfecta desigualdad (una persona tiene todos los ingresos y los demás ninguno). El índice de Gini es el coeficiente de Gini expresado en referencia a 100 como máximo, en vez de 1, y es igual al coeficiente de Gini multiplicado por 100. Una variación de dos centésimas del coeficiente de Gini (o dos unidades del índice) equivale a una distribución de un 7Aunque el coeficiente de Gini se utiliza sobre todo para medir la desigualdad en los ingresos, también puede utilizarse para medir la desigualdad en la riqueza. Este uso requiere que nadie disponga de una riqueza neta negativa.
https://es.wikipedia.org/wiki/CoeficientedeGini

3.¿Que es posit y que relacion 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, etc. El cambio de nombre de la empresa RStudio, se debe a que ha estado cambiando herramientas exclusivas de R Para mejorar su uso, y complacer las expectativas de los usuarios. El nombre que era sinónimo de desarrollo R de código abierto es un cambio de marca para representar mejor a la empresa en su conjunto