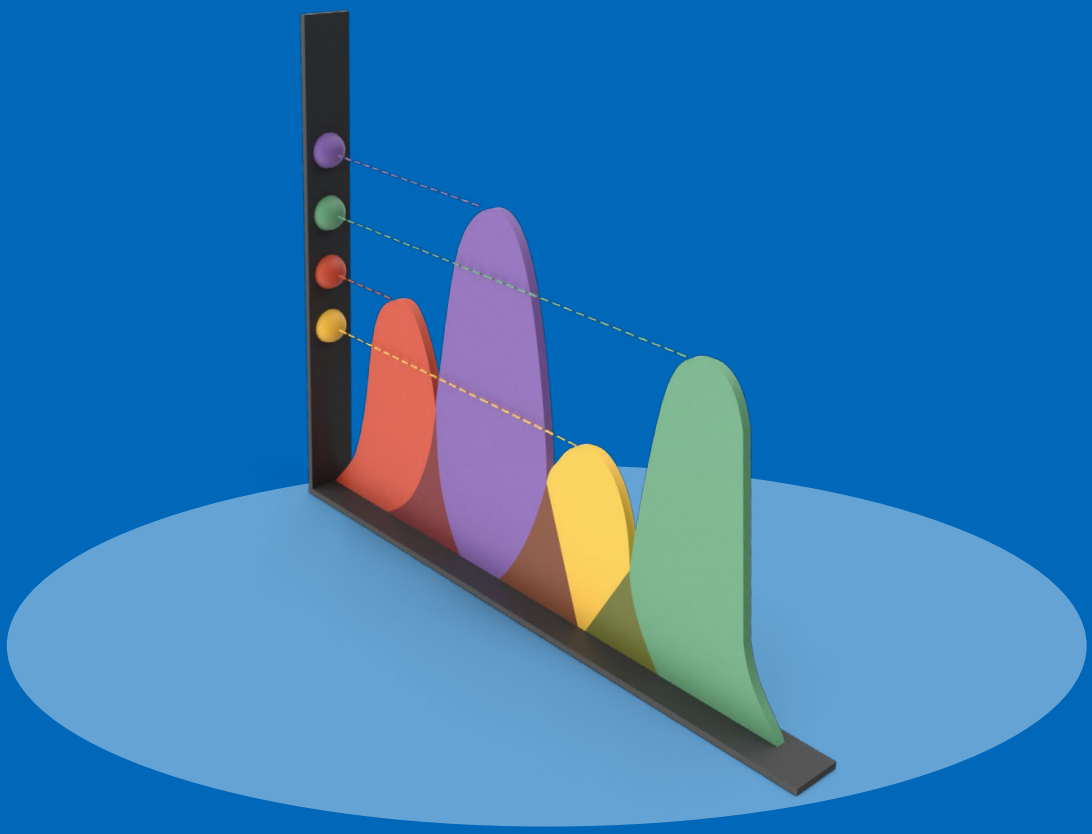
GUIDEBOOK

Introducción a las estadísticas y probabilidades para la ciencia de datos



www.dataiku.com

Introducción

A medida que pasa el tiempo y el panorama de la ciencia de datos evoluciona, también lo hacen los tipos de científicos de datos. Ahora bien, cada científico de datos tiene puntos fuertes que pueden convenir a diferentes tipos de empresas en función de los problemas o proyectos empresariales en los que estén trabajando. Mientras que algunos son generalistas (saben un poco de todo) y otros son especialistas en temas específicos o en casos prácticos, otros tienen formación de estadísticos y aplican e interpretan estadísticas en todas las áreas de la ciencia de datos. Esta guía está destinada a los científicos de datos que no tienen formación formal en estadísticas con el fin de ayudarles a entender los conceptos básicos de estadística y cuándo utilizarlos.

Aunque conocerlas no es un requisito indispensable para ser un buen científico de datos, las estadísticas son una herramienta complementaria que puede utilizarse para entender mejor los nuevos conceptos del machine learning, mejorar los modelos e incluso avisarle cuándo está haciendo algo incorrecto, lo que podría conducir a problemas mayores más adelante si no se toma en cuenta. Además, puede:

Enseñarle a pensar de otra forma y a considerar el proyecto de ciencia de datos desde otra perspectiva.

Ayudarle a formular las preguntas correctas, lo que, a su vez, le ayudará a tomar decisiones mejor fundadas.

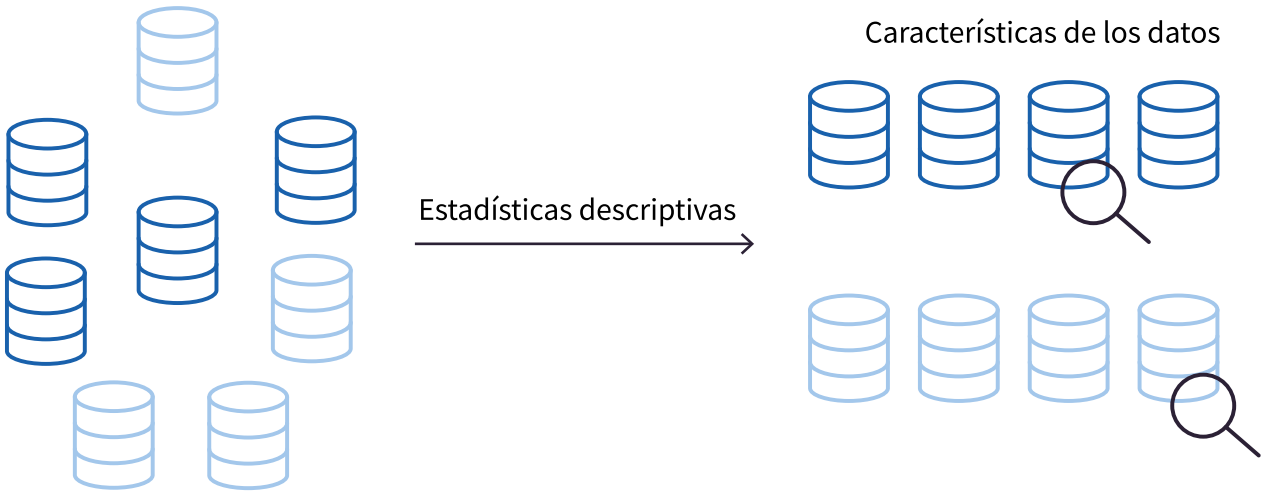
Ofrecerle un método diferente de abordar los mismos problemas, lo que le permitiría saltarse pruebas innecesarias y evitar errores.

En esta guía se abordan conceptos claves de estadística y probabilidades valiosos para los científicos de datos, cómo pueden utilizarse para mejorar aún más en el trabajo, y las funcionalidades estadísticas de Dataiku que pueden ser útiles. El conocimiento de los fundamentos estadísticos de su trabajo le permitirá aportar un nuevo punto de vista y comprensión a un proyecto de datos u obtener un resultado más rápido dominando las bases en lugar de lanzarse directamente a las llamativas técnicas del machine learning.

An Introducción a los conceptos claves de estadística y probabilidadesde

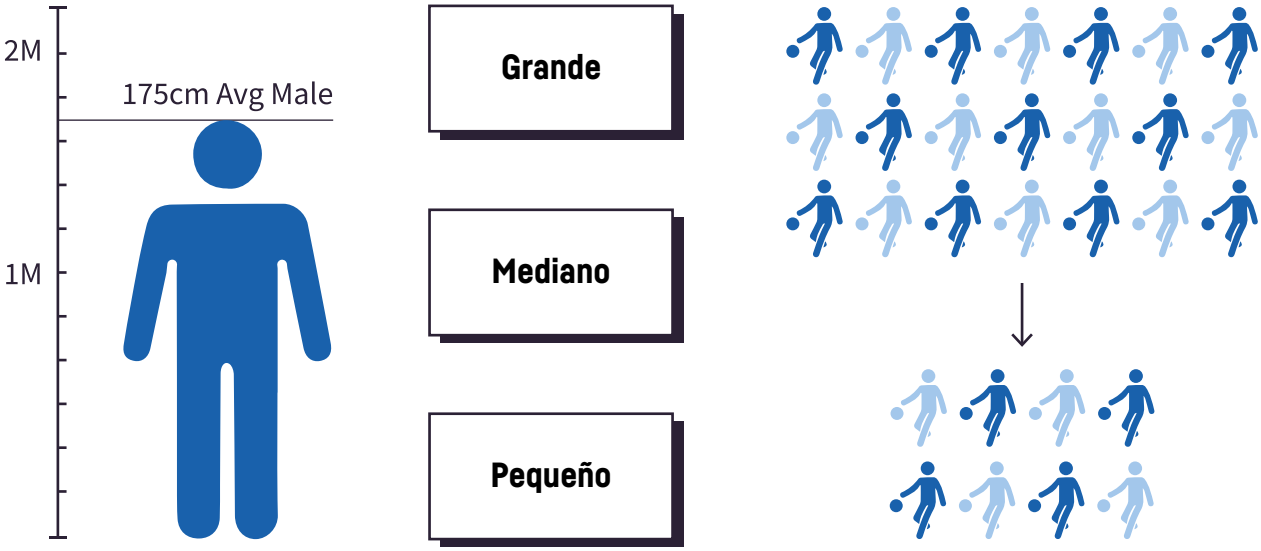
Por definición, la estadística es la ciencia que consiste en recopilar, analizar, presentar e interpretar datos[1]. Por lo tanto, la estadística puede ser una herramienta muy valiosa para los científicos de datos, dado que lo que se espera de ellos es que recopilen, limpien, preparen y analicen grandes volúmenes de datos estructurados y no estructurados y comuniquen sus hallazgos o resultados.

Existen principalmente dos categorías de estadísticas, la estadística descriptiva y la estadística inferencial. Como lo indica su nombre, la estadística descriptiva describe las características o propiedades importantes de los datos para organizarlos. Por ejemplo, si se desea conocer la estatura media de los jugadores de un equipo de baloncesto, en estadística descriptiva se registraría la estatura de cada jugador y se calcularía la estatura máxima, mínima y media de los componentes del equipo. Por lo tanto, la estadística descriptiva puede utilizarse para mostrar información resumida de los datos y ayudar a presentarlos de tal forma que tengan sentido.



[1] *https://www.britannica.com/science/statistics*

La estadística inferencial permite encontrar una propiedad en un conjunto de datos de muestra e inferir, es decir, deducir, que dicha propiedad existe en la población de la que se extrajo la muestra. Básicamente, la teoría de probabilidades puede utilizarse para sacar una conclusión satisfactoria sobre una población aunque solo se haya observado una muestra. Por ejemplo, la estatura media de los jugadores del equipo de baloncesto de la muestra podría utilizarse para inferir la estatura media de todos los jugadores de baloncesto.



Fuente: Edureka.co

Definiciones de 14 términos básicos de estadística y probabilidades

Los términos que definimos en este documento se utilizan comúnmente en estadística y es posible que conozca o reconozca algunos de ellos por su trabajo en ciencia de datos; pueden ser útiles cuando se aplican estadísticas específicamente a la ciencia de datos y al machine learning. Si trabaja en un proyecto en el que la estadística puede ser útil o si desea simplemente descubrir conceptos básicos o refrescar sus conocimientos, esperamos que estas definiciones le resulten claras y útiles.

	**Población (n)** Son las fuentes de las que deben recopilarse los datos. Presenta ciertos parámetros como la media (o promedio), la mediana, la moda, etc.		**Muestra (n)** Subconjunto aleatorio de la población. Puede utilizarse para estimar los parámetros de toda la población
	Variable (n) Cualquier característica, número o cantidad que pueda medirse o contarse.		**Parámetro (n)** Cantidad que indiza una familia de distribuciones de probabilidades (es decir, la media o la mediana de una población). Los parámetros son números que caracterizan o describen a toda una población, mientras que las estadísticas solo lo hacen para una muestra de una población.
	**Regresión (n)** Método de predicción cuyo output es un número real, es decir, un valor que representa una cantidad a lo largo de una línea.		**Probabilidad (n)** Medida numérica de la probabilidad de que ocurra un evento específico. La probabilidad utiliza una escala de 0 a 1, donde los valores cercanos a 0 indican que es improbable que ocurra un evento y los cercanos a 1 que es probable.
	**Distribución de probabilidades (n)** Función que describe los valores y probabilidades que puede tomar una variable en un rango determinado.		**Distribución muestral (n)** Distribución de probabilidades para una estadística de muestra obtenida de varias muestras extraídas de una población específica.
	**Prueba de hipótesis (n)** Método para probar la precisión con la que un modelo basado en un conjunto de datos predice la naturaleza de otros conjuntos de datos generados utilizando el mismo proceso.		**Significación estadística (n)** En las pruebas de hipótesis, se dice que un resultado tiene significación estadística, o es estadísticamente significativo, si una relación entre dos o más variables es causada por algo distinto del azar.
	**Hipótesis nula (n)** Afirmación general que establece que no existe relación entre dos fenómenos considerados o asociación entre dos grupos. Se representa con el símbolo H_0 y supone que los resultados son fruto del azar.		**Hipótesis alternativa (n)** Afirmación que establece que existe relación entre dos variables seleccionadas. Se representa con el símbolo H_1 o H_a y supone que los resultados son el fruto de causas reales.
	**Valor P (n)** Medida de la probabilidad de obtener los resultados observados cuando la hipótesis nula es cierta.		**Razonamiento bayesiano (n)** Proceso de actualización de las creencias conforme se van recopilando datos adicionales. Indica que se puede aprender de los datos faltantes o incompletos y de las aproximaciones.

** Entre las fuentes en inglés consultadas para estas definiciones cabe destacar Britannica, Investopedia, KDNuggets y Towards Data Science.*

Conceptos básicos de estadística útiles para cualquier científico de datos

Prueba de hipótesis

Veamos más en detalle uno de los conceptos de la tabla anterior: la prueba o test de hipótesis. Muchos de los fundamentos básicos de estadística pueden resumirse a una prueba de hipótesis. En esencia, las estadísticas –y los científicos de datos formados en estadísticas– intentan encontrar respuesta a interrogantes muy sencillos a los que es extremadamente difícil responder, como por ejemplo "¿hay alguna diferencia entre el Grupo A y el Grupo B?"

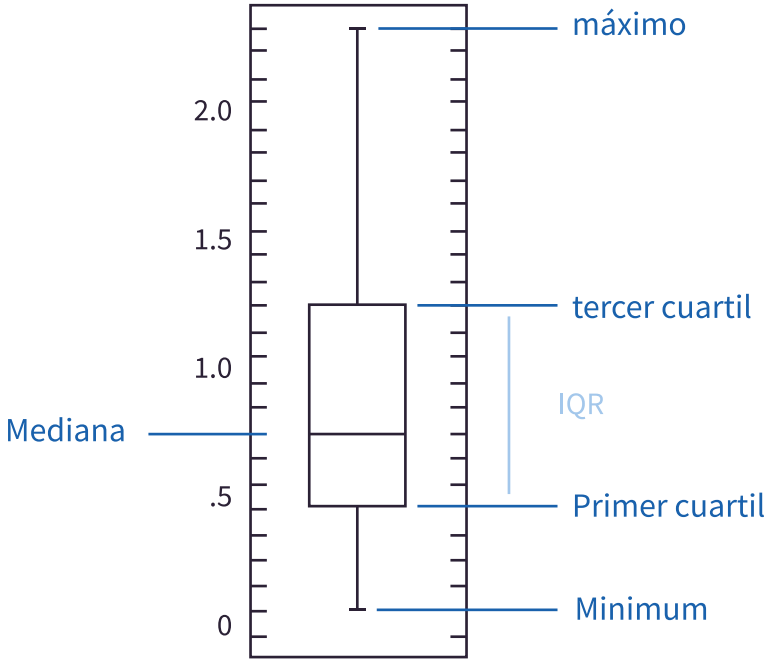
Las pruebas de hipótesis utilizan datos de una muestra para sacar conclusiones sobre un parámetro de población o la distribución de probabilidades de la población. En primer lugar, se establece un supuesto preliminar relativo al parámetro o la distribución, que se conoce como hipótesis nula. Luego se define lo contrario de lo que se afirma en la hipótesis nula, que se conoce como hipótesis alternativa (ambas se tratan en la sección de definiciones más arriba).

A continuación, se utilizan datos de muestra para determinar si la hipótesis nula puede rechazarse. Si es el caso, la conclusión estadística es que la hipótesis alternativa tiene mayores probabilidades de ser verdadera que la hipótesis nula. Las pruebas de hipótesis básicas también pueden ayudar a ahorrar tiempo, ya que permiten hacer una pregunta y probarla antes de emprender el desarrollo de un complicado modelo de machine learning.

Pongamos por ejemplo que dos ciudades vecinas, Eastchester y Westchester, realizan el mismo examen estandarizado a sus estudiantes. Cabría preguntarse: "¿Difiere la nota o calificación media obtenida en el examen por los estudiantes de Eastchester y Westchester?" La hipótesis nula sería: "No, la nota media del examen es la misma entre los estudiantes de Eastchester y Westchester." La afirmación opuesta, es decir, la hipótesis alternativa, sería: "Sí, la nota media del examen es diferente entre los estudiantes de Eastchester y Westchester." Para realizar esta prueba de hipótesis, se puede recopilar una muestra aleatoria de las notas de los exámenes de los estudiantes de ambas ciudades y calcular un valor p para orientar las conclusiones.

Características estadísticas

Las características estadísticas, un concepto estadístico popular en ciencia de datos, entran en juego durante la fase de exploración de los datos y engloban temas como sesgo, varianza, media, mediana y percentiles. En el diagrama de caja básico que figura a continuación, los valores mínimo y máximo representan el umbral superior e inferior del rango de datos. El "primer cuartil" se utiliza para demostrar que el 25 % de los puntos de datos están por debajo de dicho valor, mientras que el "tercer cuartil" muestra que el 75 % de los puntos de datos están por debajo de dicho valor.



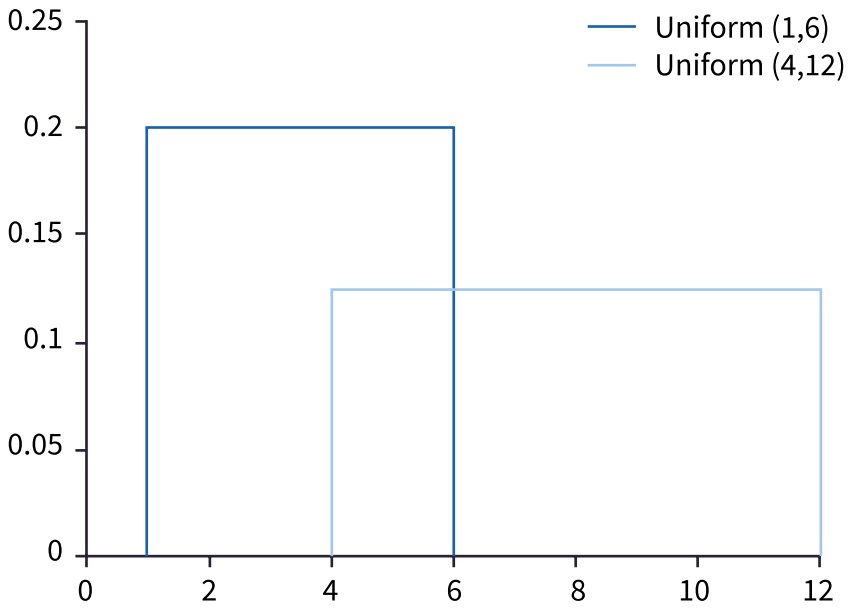
Fuente: Medium[2]

El diagrama de caja puede utilizarse para interpretar correctamente los resultados de los datos. Por ejemplo, si el diagrama de caja es pequeño, muchos de los puntos de datos tendrán valores similares. Si es grande, puede indicar que los puntos de datos son diversos, ya que los valores se distribuyen en un amplio rango. Si el valor mediano está más cerca de la parte inferior que de la superior, podemos deducir que los datos tienen generalmente valores más bajos y menos valores altos, mientras que si está más cerca de la parte superior, sabemos que los datos tienen generalmente valores más altos con menos valores bajos. Si la mediana no cae en el medio del diagrama, podría ser un indicador de que los datos están sesgados y deben estudiarse más de cerca antes de utilizarlos en un modelo.

[2] *https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae*

Distribución de probabilidades

Una probabilidad es la posibilidad de que un evento ocurra aleatoriamente. En ciencia de datos, se suele cuantificar en el rango de 0 a 1, donde 0 significa que el evento no ocurrirá y 1 indica la certeza de que sí ocurrirá. Cuanto mayor sea la probabilidad de un evento, mayor será la posibilidad de que ocurra realmente. Por lo tanto, una distribución de probabilidades es una función que representa la probabilidad de obtener los valores que puede tomar una variable aleatoria. Se utilizan para indicar la probabilidad de que ocurra un evento o de obtener cierto resultado.



Fuente: Medium[3]

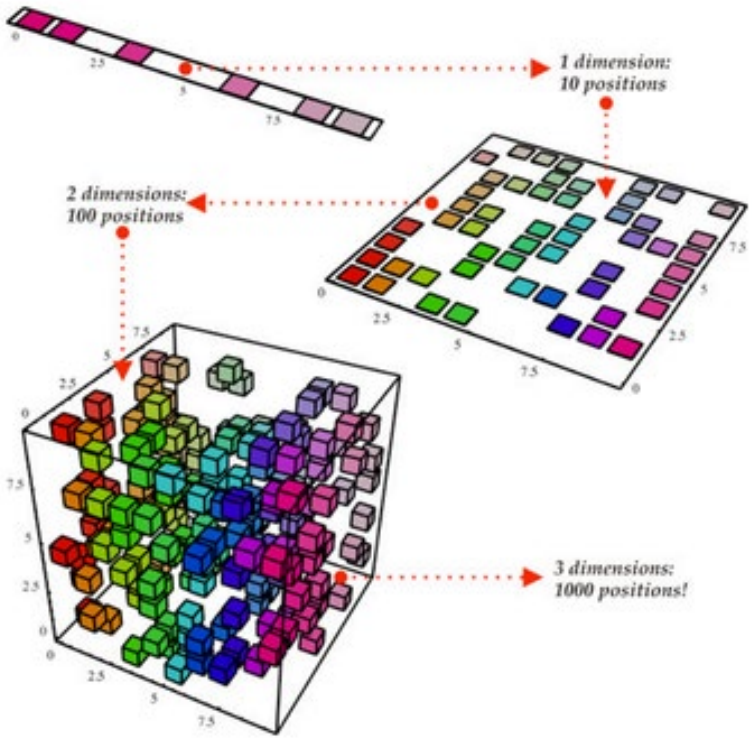
La imagen anterior ilustra la distribución de probabilidades más básica, una distribución uniforme. Esto indica que todos los valores en un rango dado son igualmente probables, mientras que los demás nunca ocurrirán.

[3] *https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae*

Reducción de la dimensionalidad

La reducción de la dimensionalidad en el ámbito de la ciencia de datos consiste en reducir el número de dimensiones de un conjunto de datos, es decir, el número de variables. Si tomamos un cubo para representar un conjunto de datos tridimensional con 1000 puntos de datos y lo proyectamos en un plano bidimensional, solo vemos una de sus caras lo que mejora la visualización, ya que es más fácil visualizar un gráfico en un espacio bidimensional que en uno tridimensional y con algoritmos de machine learning se requieren menos recursos para aprender en espacios bidimensionales que en tridimensionales.

La reducción de la dimensionalidad también puede realizarse mediante la poda o "pruning", que permite suprimir características que no serán importantes en el análisis final. Por ejemplo, si después de revisar un conjunto de datos nos damos cuenta de que 14 de 20 características tienen una alta correlación con el output, pero las otras seis tienen una baja correlación, lo más sensato sería suprimirlas del análisis en esa etapa sin que haya repercusiones negativas en el output.



Fuente: Medium[4]

[4] *https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae*

Razonamiento bayesiano

El razonamiento bayesiano requiere utilizar las probabilidades para modelizar los procesos de muestreo y cuantificar la incertidumbre antes de recopilar los datos. Este nivel de incertidumbre antes de recopilar los datos se conoce como "probabilidad a priori", y se actualiza a "probabilidad a posteriori" después de recopilar los datos. La International Society for Bayesian Analysis explica el Teorema de Bayes de la siguiente forma: "En el paradigma bayesiano, el conocimiento inicial de los parámetros del modelo se expresa mediante una distribución de probabilidades en los parámetros, llamada distribución a priori".[5]

La distribución a priori es el conocimiento inicial de un tema, y cuando se descubre nueva información, se expresa como "probabilidad". Esto pone de relieve que un conocimiento básico del razonamiento bayesiano puede resultar útil cuando se trabaja con modelos de machine learning más complejos.

[5] *https://bayesian.org/what-is-bayesian-analysis/*

Aplicación de las estadísticas a la ciencia de datos

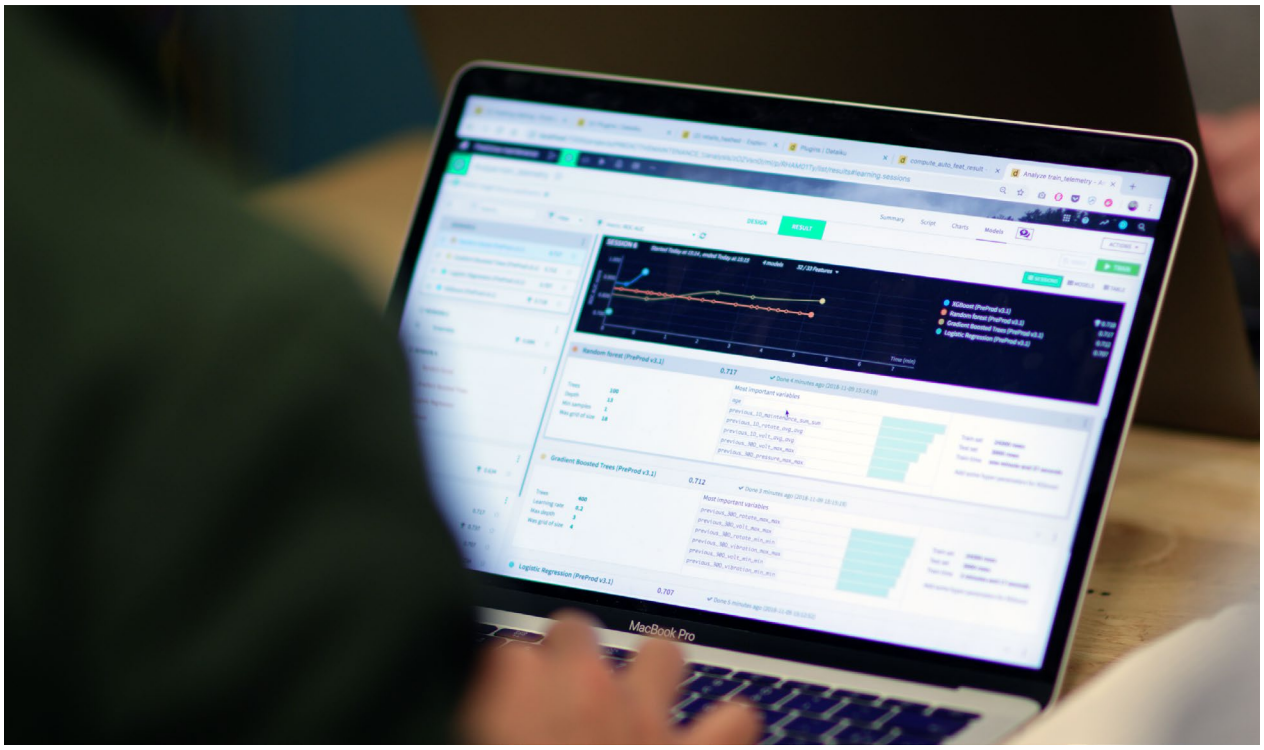
Las estadísticas y la ciencia de datos están muy relacionadas pero, a la postre, son disciplinas muy distintas. Ahora bien, los científicos de datos utilizan conjuntos de datos de ambas disciplinas para sacar conclusiones y sus propias observaciones sobre el mundo. Intentar entender la correlación entre los datos de entrada y de salida forma parte de las estadísticas, mientras que la recopilación de datos, el diseño de experimentos basados en datos, y la aplicación de las estadísticas y el machine learning para entender el significado de dichos datos forma parte

de la ciencia de datos. Las estadísticas y la ciencia de datos están muy relacionadas pero, a la postre, son disciplinas muy distintas. Ahora bien, los científicos de datos utilizan conjuntos de datos de ambas disciplinas para sacar conclusiones y sus propias observaciones sobre el mundo. Intentar entender la correlación entre los datos de entrada y de salida forma parte de las estadísticas, mientras que la recopilación de datos, el diseño de experimentos basados en datos, y la aplicación de las estadísticas y el machine learning para entender el significado de dichos datos forma parte de la ciencia de datos.

Aunque las preguntas que se plantean en ambas disciplinas no son de la misma naturaleza, las estadísticas pueden utilizarse antes de abordar cualquier problema de ciencia de datos. Muchos se lanzan directamente al machine learning sin pasar por las etapas básicas previas necesarias. El principio de Pareto se aplica en este caso, ya que en general el 80 % del resultado puede obtenerse con el 20 % de las herramientas. Por lo tanto, unas buenas bases estadísticas pueden contribuir a simplificarle el trabajo.

Supongamos que está probando varios modelos de machine learning para un caso específico de un cliente. Algunos de los modelos asumen distribuciones de probabilidades específicas (los valores y probabilidades que puede tomar una variable en un rango dado) de datos de entrada. Por lo tanto, dado su trabajo, debe ser capaz de identificarlas y ajustar los datos de entrada consecuentemente. Este es solo un pequeño ejemplo de cómo las estadísticas pueden integrarse en el día a día de la ciencia de datos.

Conocer las estadísticas le ayudará a determinar la diferencia entre los resultados que son creíbles y los que probablemente ocurrieron al azar. Todos los proyectos de ciencia de datos implican una pequeña dosis de análisis exploratorio de datos (EDA) para conocer mejor los datos con los que se va a trabajar (es decir, resumir o describir muestras de datos tanto numérica como visualmente).

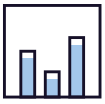


A menudo, los científicos de datos (y sus clientes) dirán que disponen de un gran volumen de datos, pero que no están seguros de las preguntas que deben formular o por dónde empezar a extraer valor de los datos. Las estadísticas pueden ayudar a sentar las bases y a identificar patrones e insights. En cierto modo, los modelos de machine learning suelen hacer preguntas estadísticas simples de los datos varias veces, y agregan las respuestas para hacer una predicción o descubrir la estructura general. En la siguiente sección veremos cómo Dataiku pone a disposición de los científicos de datos una herramienta específica para EDA.

Por otra parte, los científicos de datos que utilizan modelos de machine learning y sistemas de IA se enfrentan hoy al serio desafío de equilibrar la interpretabilidad y la precisión derivadas de la diferencia entre los modelos de caja negra y de caja blanca. Mediante la IA y el machine learning, deben determinar con frecuencia el grado de precisión o lo que significa la predicción (cuyo resultado es un número o un valor), mientras que con las estadísticas interpretan los resultados para intentar describir de dónde sale tal etiqueta. Las estadísticas ofrecen una capa adicional de interpretación, ya que permiten reforzar las razones que llevan a creer en un resultado específico.

Qué estadísticas y probabilidades pueden ser útiles para los científicos de datos (y cómo puede ayudarles Dataiku)

Como dijimos anteriormente, las estadísticas y las probabilidades son un valor añadido para los científicos de datos, ya que ofrecen un nuevo enfoque y una nueva perspectiva antes de sumergirse en un proyecto de machine learning. Con Dataiku, los científicos de datos (con o sin formación en estadísticas) pueden realizar análisis estadísticos avanzados en hojas de cálculo y fichas, a la vez que estrechan la colaboración con el equipo de datos o análisis. Esta hoja de cálculo ofrece una interfaz específica para realizar tareas de EDA, que permite:



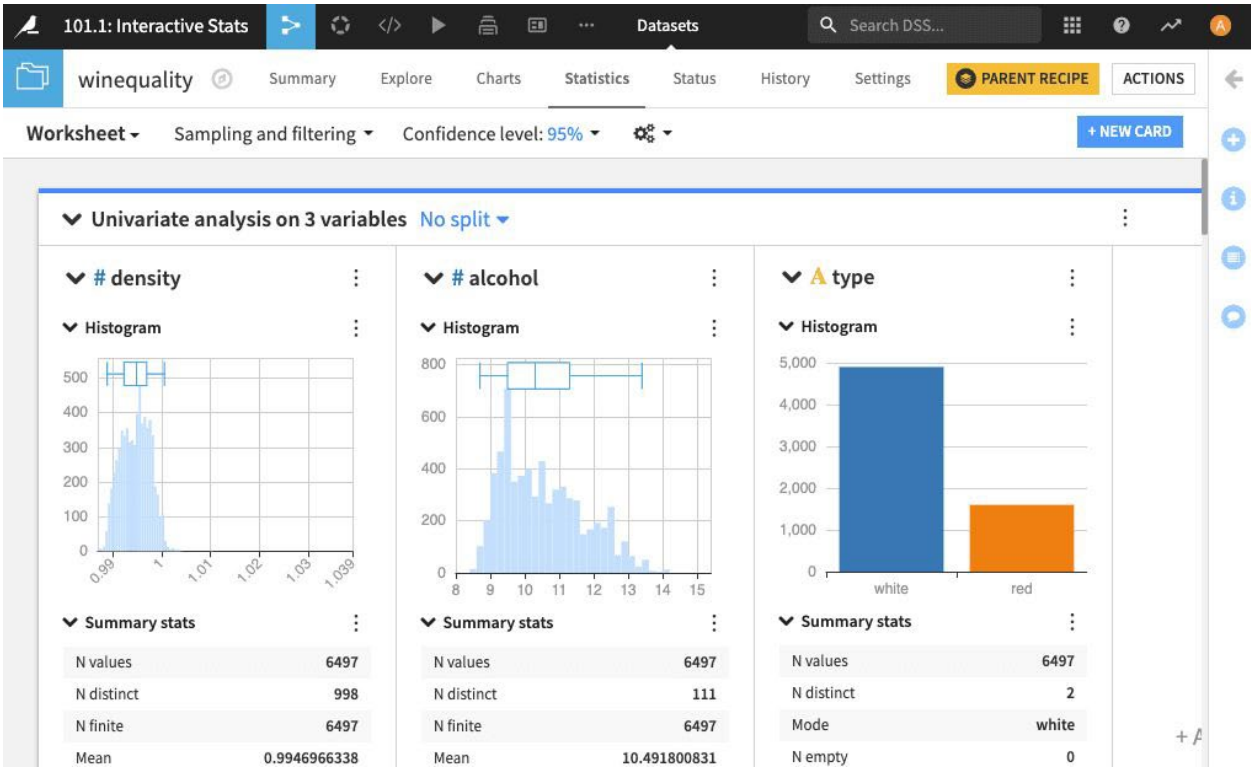
Resumir o describir muestras de datos (es decir, mediante análisis univariable, análisis bivariable, distribución y ajuste de curvas, y matrices de correlación), lo que forma parte de la estadística descriptiva

Extraer conclusiones de un conjunto de datos de muestra sobre una población subyacente (es decir, mediante pruebas de hipótesis), lo que forma parte de la estadística inferencial.

Visualizar la estructura del conjunto de datos en un número reducido de dimensiones, utilizando el análisis de componentes principales (ACP), lo que forma parte de la reducción de la dimensionalidad.

Por otra parte, las fichas de una hoja de cálculo permiten realizar directamente varias tareas estadísticas a la vez y mantener el espacio de trabajo bien organizado. En Dataiku, una ficha se utiliza para realizar una tarea EDA específica, como describir el conjunto de datos, extraer inferencias sobre una población subyacente o analizar el efecto de la reducción de la dimensionalidad.

Una hoja de cálculo puede tener varias fichas, por lo que cuando se crea una nueva, hay que especificar el tipo y los valores de parámetros correspondientes. El conjunto de datos también puede agruparse por una variable específica (para poder realizar cálculos en cada subgrupo de datos y comparar estadísticas entre varios grupos) utilizando el menú "Split by" en una ficha.



Tipos de Fichas Estadísticas en Dataiku

1. Análisis univariable

Este enfoque permite explorar conjuntos de datos de una variable simultáneamente. No tiene en cuenta las relaciones entre dos o más variables de un conjunto de datos y ayuda a describir y resumir el conjunto de datos mediante una variable. Permite seleccionar múltiples variables del conjunto de datos para visualizar las distribuciones individuales de las variables una al lado de la otra. Dependiendo del tipo de variable, Dataiku llena cada sección de la ficha con el análisis estadístico apropiado (es decir, histograma, diagrama de caja, estadísticas compendiadas, tabla de cuantiles, tabla de frecuencias).

2. Análisis bivariable

A diferencia del análisis univariable, el análisis bivariable permite analizar dos variables y determinar si existe una relación entre ambas (donde una variable es la variable de respuesta y la otra una variable de factor). La asignación de funciones a variables en el conjunto de datos determinará cómo las utiliza Dataiku en fichas estadísticas. Dependiendo de los tipos de variables de factor y respuesta, Dataiku llena cada sección de la ficha del mismo modo que en un análisis univariable (es decir, gráfico de mosaico, diagrama de dispersión, diagrama de caja).

3. Ajuste de curvas y distribuciones

Las fichas que se describen a continuación modelizan las distribuciones o relaciones de variables numéricas:

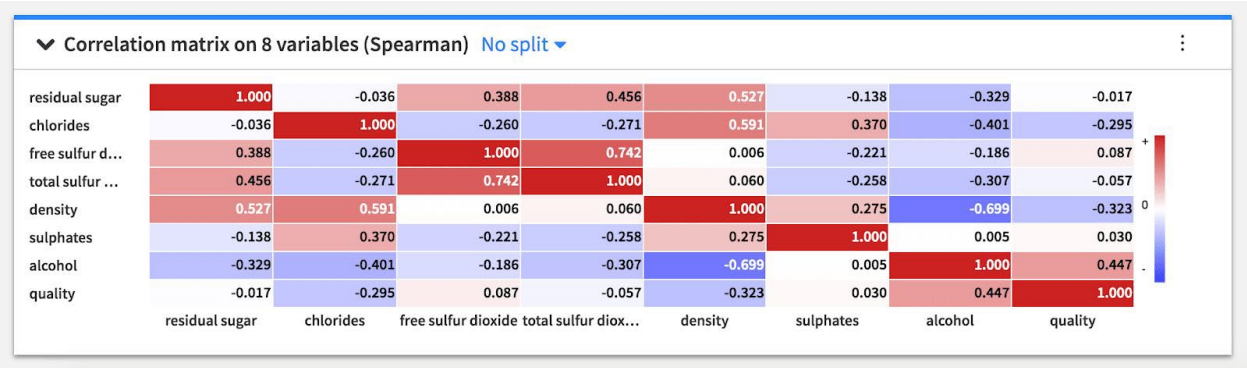
Fit distribution: Calcula los parámetros de las distribuciones de probabilidad de una variable especificada en el conjunto de datos

2D Fit distribution: Visualiza la densidad de las distribuciones bivariables esquematizando la estimación de densidad de Kernel (KDE) o la distribución normal (gaussiana)

Fit curve: Modeliza la relación entre dos variables creando una o más ajustes de curvas

4. Matriz de correlación

Una matriz de correlación puede utilizarse para mostrar los coeficientes de correlación (o grado de relación) entre variables, especialmente cuando un conjunto de datos presenta un gran número de columnas. La matriz de correlación es simétrica, ya que la correlación entre una variable V_1 y una variable V_2 es la misma que la correlación entre la variable V2 y la V1. Además, los valores en la diagonal son siempre iguales a uno, porque una variable siempre está perfectamente correlacionada consigo misma. En Dataiku, la ficha matriz de correlación permite ver una tabla visual de las correlaciones por pares de múltiples variables de un conjunto de datos, como a continuación:



5. Pruebas estadísticas

Dataiku contiene una infinidad de fichas de pruebas estadísticas, agrupadas en pruebas de una muestra, pruebas de dos muestras, pruebas de N muestras, pruebas de localización o distribución, y pruebas categoriales. Aunque el objetivo y alcance de cada una de ellas es diferente, estas pruebas permiten tomar decisiones cuantitativas probando hipótesis estadísticas. Cada ficha muestra el resultado de una prueba estadística y permite obtener más información sobre la prueba (lo que hace, supuestos de base, etc.) directamente en el encabezado de la ficha.

6. PCA

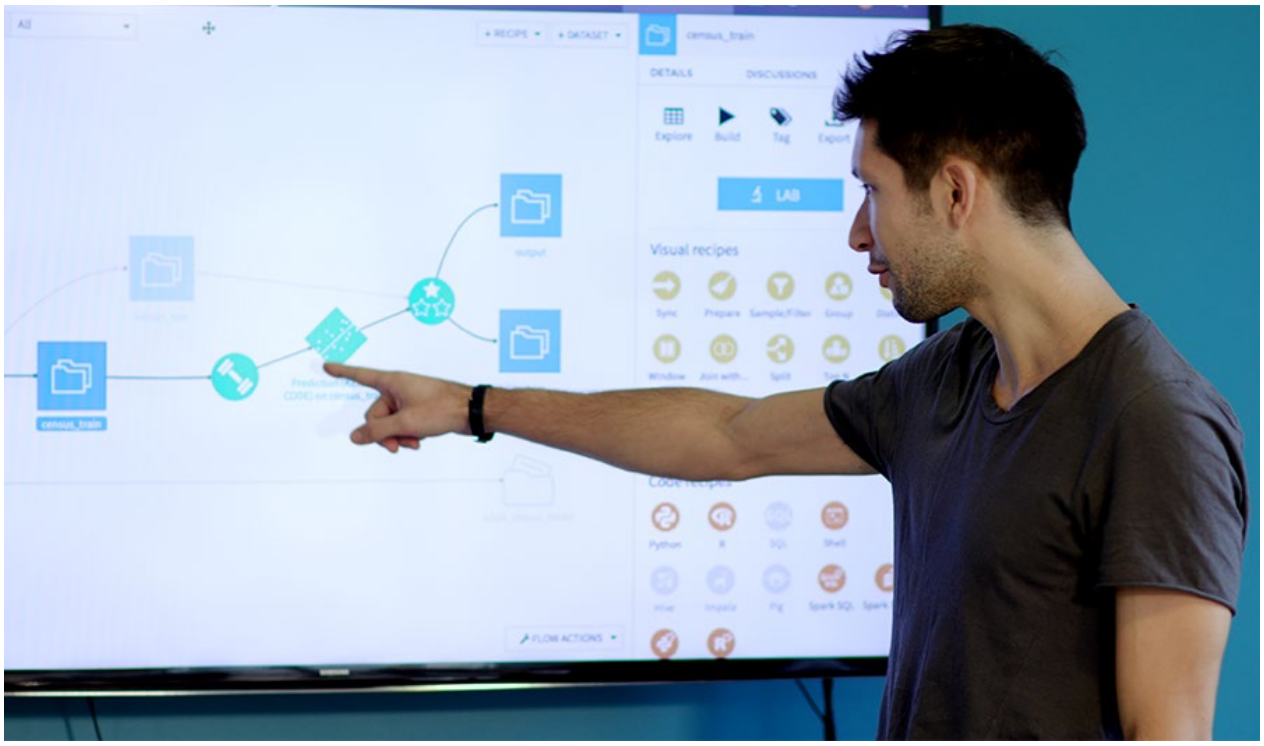
PCA es sumamente útil para reducir la dimensionalidad de un conjunto de datos. Tácticamente, realiza una transformación lineal de un conjunto de datos (que probablemente contiene variables correlacionadas) en una dimensión de variables no correlacionadas linealmente (denominadas componentes principales). El objetivo de esta etapa es maximizar la varianza de los datos en el menor número de dimensiones. En Dataiku, la ficha PCA permite obtener una representación visual de un conjunto de datos en una dimensión reducida.

Conclusión

La demanda de científicos de datos no va a declinar por el momento. Por lo tanto, aunque no tenga un diploma o una formación en estadísticas y probabilidades no es demasiado tarde para aprender; solo tendrá que ponerle dedicación y empeño.

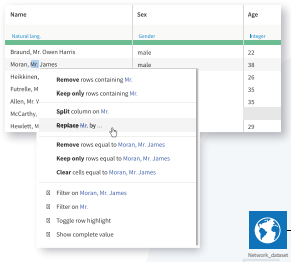
Familiarizarse con los fundamentos básicos del análisis estadístico y las probabilidades le ofrecerá una ventaja competitiva. Dado que una parte significativa de los proyectos de ciencia de datos y machine learning se basa en el análisis de datos, conocer estos conceptos le permitirá generar mejores insights y tomar decisiones más informadas gracias a los conjuntos de datos.

Tener conocimientos en estadísticas y probabilidades no solo le ayudará a progresar en su carrera de científico de datos, sino que además puede ayudarle a ponerle más sabor a su trabajo y entenderlo mejor puesto que sabe cuáles los cimientos en los que se basa el machine learning. También puede ayudarle a mejorar sus modelos de machine learning, ya que entenderá mejor sus mecanismos (lo que, de por sí, contribuye a entender mejor los modelos de caja negra). Esperamos que esta guía le haya resultado útil en su viaje hacia la integración del razonamiento estadístico en sus proyectos de ciencia de datos.

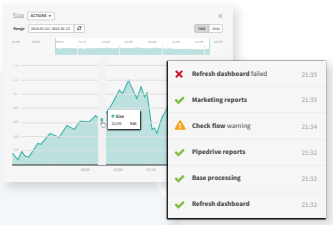


Your Path to Enterprise AI

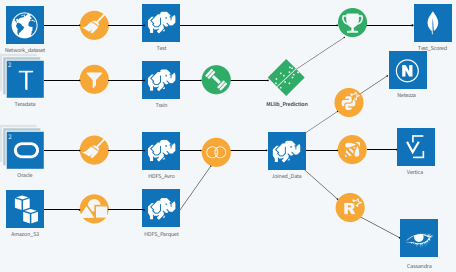
Limpieza & Preparación

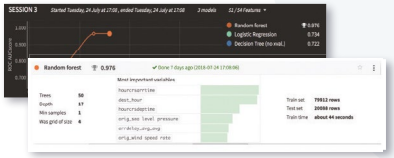


Monitoreo y Ajustes

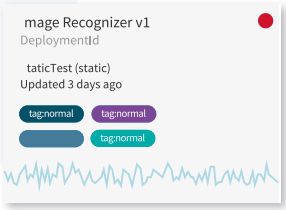


Construcción + Aplicación del Machine Learning

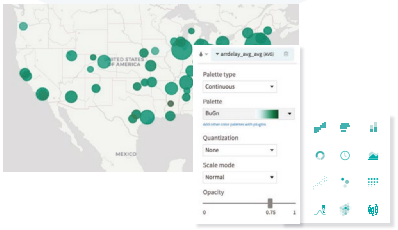




Despliegue en producción



Extracción & Visualisation



400+
CLIENTES

40,000+
USUARIOS ACTIVOS*

*científicos de datos, analistas, ingenieros y demás

Dataiku es una de las plataformas de IA y machine learning líderes del sector a escala mundial, que apoya las iniciativas de datos de las empresa gracias a un sistema de IA colaborativo, flexible y responsable, a escala de la empresa. Cientos de empresas utilizan Dataiku para respaldar sus operaciones y asegurarse de que siguen siendo relevantes en un mundo en constante evolución.

GUIDEBOOK

www.dataiku.com