FEB 11

SPRINT 1 - ENCUENTRO 4

Probabilidad y Estadística - Pandas

Grabación

Presentación

El equipo docente subió contenido extra del encuentro

Lista de recursos

Comunidad

Notebook: DS_Bitácora_04_Estadísti

Contenido extra

DS-ONLINE-65 2/2/21

Volver a las bases

> *"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." - John Wills,* *Creador de Crunch.*

Unas pizcas de matemática

¿Cuáles son los ingredientes de un/a Data Scientist? Data Science tiene mucho de programación pero, ¿programar para qué?. En primer lugar, programamos para abrir nuestras bases de datos y así, una vez que contamos con nuestros datos, poder analizarlos y transformarlos. Esto que hacemos está guiado, principalmente, por dos (¿una?) disciplinas que seguramente conoces: Probabilidad y Estadística. Por lo tanto, para empezar a orientarnos en Data Science, es necesario que repasemos estos conceptos clave que, quizás, dejamos olvidados en la memoria una vez que terminamos la escuela.

Vamos a plantearte una situación y, antes de seguir leyendo, es fundamental que reflexiones sobre ella:

Vamos al casino a jugar Craps (un juego de azar que consiste en apostar en base a lo que esperamos que salga de una tirada de dados) pero comenzamos a sospechar que uno de los dados que estamos utilizando está "cargado", ¿cómo podremos saber cuál?

Una forma de averiguarlo sería hacer un análisis físico de los dados y buscar desbalances en su construcción que favorezcan una cara por sobre otra. Otra opción, la estadística, es hacer muchas tiradas de los dados, anotar sus resultados y ver si encontramos algo distinto de lo esperado. Pero esto genera muchas nuevas preguntas:

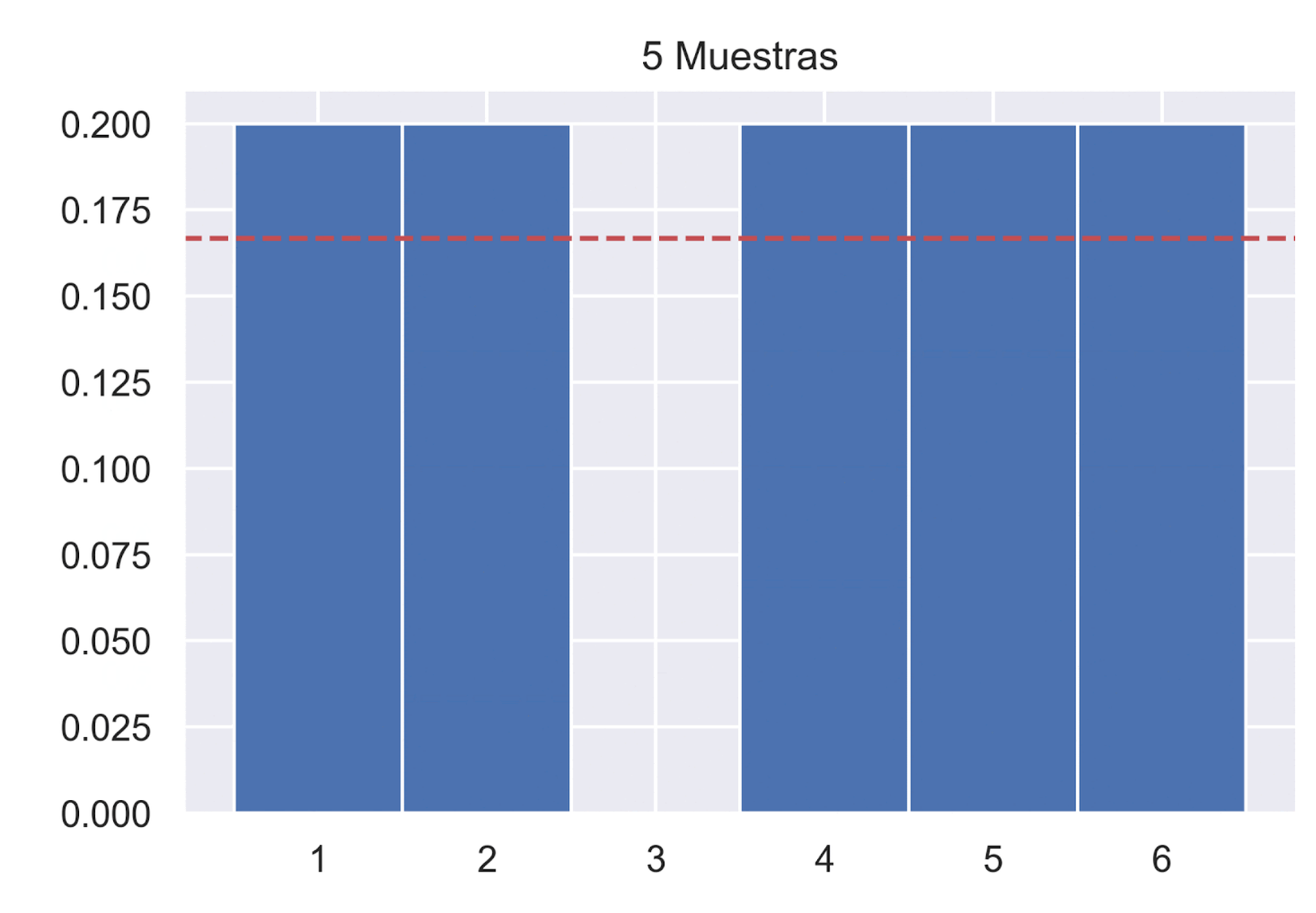
- ¿Qué es lo esperado? ¿Cómo deberían ser los resultados de tirar un dado no-cargado?
- ¿Qué significa que un dado está cargado? ¿Cómo esperamos que sean los resultados de tirar un dado cargado?
- ¿Cuándo podremos saber que algo es distinto de lo esperado? Es decir, ¿cuántas veces deberemos tirar los dados para estar seguros de cuál está cargado (si es que lo estaba)? ¿Dependerá esta decisión de los resultados obtenidos?

Para reflexionar sobre estas preguntas, no es indispensable ser técnico/a o o usar matemáticas: lo importante es que nos concentremos sobre el problema. Una buena opción para hacerlo son los ejemplos: ¿qué confianza tienes en las siguientes afirmaciones?

1. Tiré dos veces cada dado y, en uno de los dados, ambas veces salió 6 (seis), por lo que está cargado para favorecer ese número.
2. Tiré 100 veces cada dado y, en uno de los dados, todas las veces salió 6 (seis), por lo que está cargado para favorecer ese número.
3. Tiré 100 veces cada dado y, en uno de los 20 veces salió 6 (seis), mientras que en el otro solamente 12, por lo que el primer dado está cargado para favorecer ese número.

Tip: siempre es una buena práctica anotar nuestras reflexiones para repasarlas en el futuro.

En ésta y la próxima bitácora, haremos un repaso de los principales conceptos de Probabilidad y Estadística, y de la relación entre ellas. A lo largo de las entradas, retomaremos este ejemplo, por lo que podrás aprovechar mucho más las explicaciones si ya estuviste analizándolo.



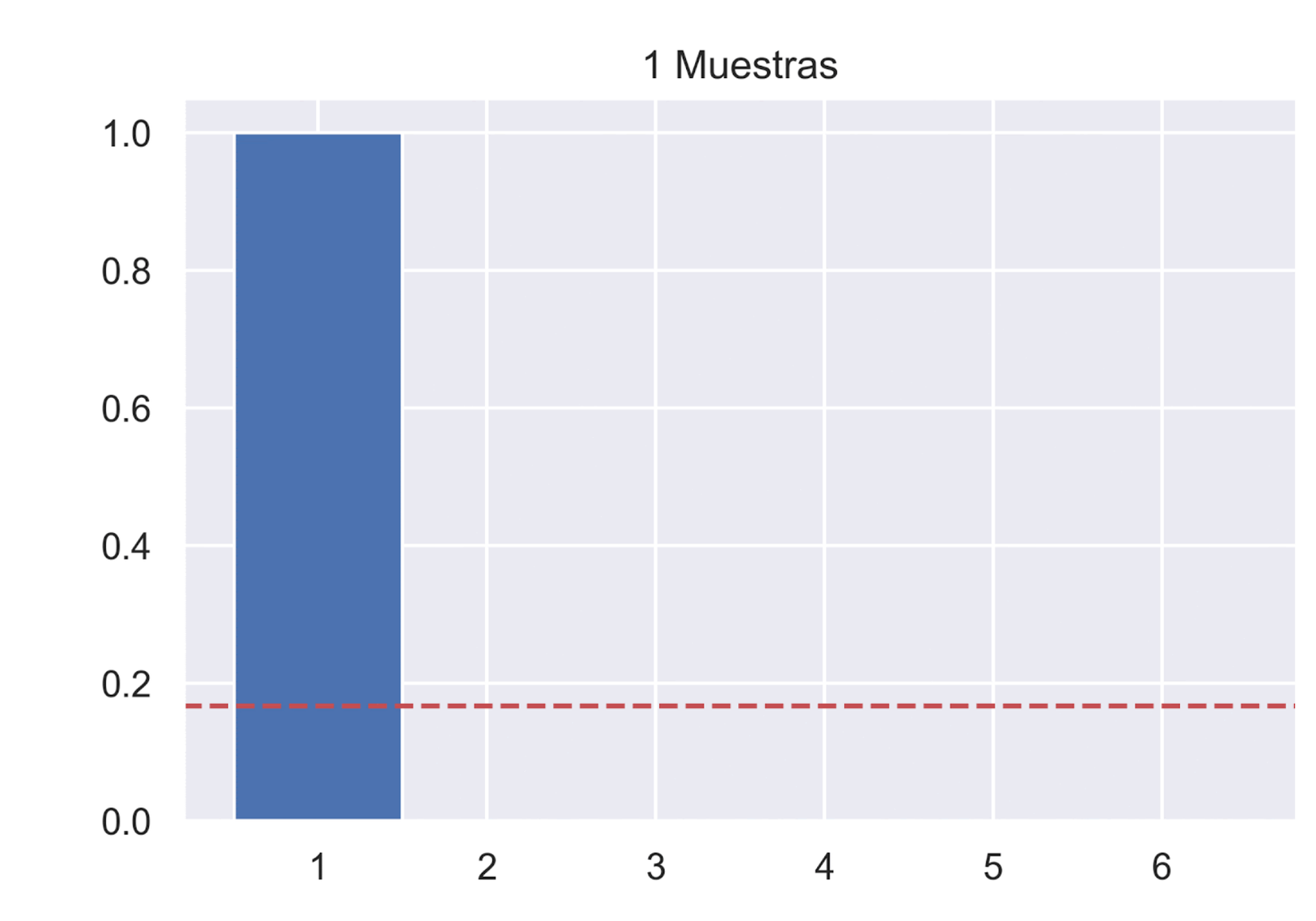


Figura 1: tomamos dos dados y anotamos los resultados de tirar cada uno. Luego, graficamos el porcentaje de veces que cae de cada lado. La línea punteada corresponde a lo esperado para un dado perfectamente balanceado. ¿Cuál será el que está cargado?

Estadística descriptiva

En el ejemplo anterior, al tirar los dados varias veces, comenzamos a tener muchos datos observados, por lo que necesitamos, por un lado, una forma de estructurarlos y, por otro, un lenguaje común para hablar de ellos. La primera pregunta que nos surge entonces es: ¿cómo vienen los datos?

En ocasiones —si tenemos suerte—, los datos vienen en formatos ordenados y prolijos como tablas. Estos se conocen como **datasets estructurados,** similar a una planilla de cálculo de Excel. En general, se suele tomar lo siguiente como válido:

- cada columna corresponde a una característica,
- cada fila corresponde a una instancia.

Por ejemplo, en el dataset de un médico cardiólogo podríamos encontrar muchas columnas: una que corresponde a la edad, una a la altura, otra al peso, género, a la presión arterial, al índice de colesterol, enfermedades preexistentes, una con el diagnóstico, etc., mientras que cada fila corresponde a un paciente. Algunos formatos en los que encontramos datasets estructurados son .CSV (comma separated value), .txt, .xlsx, .json, en bases de datos SQL, entre otros.

Sin embargo, también podemos encontrar datos en lo que se conoce como **datasets no estructurados.** Pueden ser un conjunto de audios, imágenes o hasta un corpus de texto en crudo. En general, los datasets no estructurados son difíciles de comprender para las computadoras, pero fácil para los humanos. Lograr que una computadora reconozca el contenido de una imagen o un audio es difícil, mientras que para un humano suele ser inmediato. Una de las técnicas para trabajar con datasets no estructurados es convertirlos en datasets estructurados, como por ejemplo en los N-gramas de Google.

La mayoría de las herramientas que veremos durante la carrera sirven para trabajar con datasets estructurados.

Tipos de datos

Volvamos al ejemplo del dataset del médico. Viendo la tabla, notamos que no todas las columnas tienen el mismo tipo de información, a saber:

- Las que contienen datos como la edad, altura y peso son c**olumnas cuantitativas o numéricas.**
- Las que contienen etiquetas como el género, enfermedades preexistentes o el diagnóstico con **variables cualitativas o categóricas.**

Profundizaremos mucho más durante el segundo bloque de la carrera en la taxonomía de los tipos de datos. Pero así como en programación, cuando trabajemos con un dataset es muy importante prestarle atención al tipo de dato con el que estamos trabajando, ya que todos contienen información de la cual podemos aprender.

Primeras herramientas

Ya tenemos nuestro dataset. ¿Qué podemos hacer con él?

Generalmente una de las primeras herramientas que aplicamos cuando comenzamos a trabajar con un conjunto de datos es la **estadística descriptiva.** Ésta nos da la oportunidad de encontrar propiedades en un conjunto o muestra de datos, lo que nos permite resumir los datos observados y nos da un lenguaje común para hablar sobre ellos.

- En el caso de las variables categóricas, podemos contar cuántas categorías hay y cuántas instancias pertenecen a cada categoría. Por ejemplo, en el dataset del médico, podríamos contar cuántos diagnósticos distintos hay y cuántos pacientes recibieron el mismo diagnóstico.
- En el caso de las variables numéricas, podemos hacer algunas cosas más. En particular, generar valores estadísticos. Un valor estadístico es una medida cuantitativa derivada de un conjunto de datos. Dicho así parece complicado, pero solemos hacerlo todo el tiempo; por ejemplo, el promedio para resumir un conjunto de calificaciones obtenidos en nuestros exámenes. Lo que hacemos es representar o resumir un conjunto de datos que por lo general es grande, por un nuevo conjunto de valores, en general pequeño. En este caso utilizamos el promedio, pero también podría ser su máximo, mínimo, y algunas medidas nuevas que veremos a continuación. Así, es más fácil comparar, hablar o aprender de nuestros datos que si tuviéramos que trabajar con todos ellos a la vez.

Hay dos principales tipos de valores estadísticos, que nos dan mucha información sobre cómo es la distribución de nuestros datos:

1. **Medidas de tendencia central:** que muestran el centro alrededor del cual se distribuyen los datos y son muy útiles para resumir un conjunto de datos en un solo valor. Las más conocidas son:
   - **Media** es el valor promedio estándar. Por ejemplo, si las notas de cinco exámenes son 7, 9, 8, 7 y 10, su promedio es (7 + 9 + 8 + 7 + 10)/5 = 8.2
   - **Mediana** es el valor medio exacto en un conjunto de datos ordenados. Es decir, el 50% de los valores son menores que la media y el 50% son mayores. En el mismo ejemplo, las notas de los exámenes ordenadas son 7, 7, 8, 9, 10, por lo que su mediana es 8.
   - **Moda** el valor con mayor frecuencia en un conjunto de datos. En el ejemplo es, 7, ya que aparece dos veces frente al resto de los valores que solamente aparece una.

Media, mediana y moda no coinciden. Según el problema en el que estemos trabajando, a veces una u otra serán más representativas de nuestro conjunto de datos. Existen otras medidas que iremos viendo más adelante.

1. **Medidas de variación o variabilidad:** que muestran la dispersión de datos, es decir, qué tan lejos se encuentran las mediciones del centro, y sirven para cuantificar cuán dispersos están los datos con respecto a su centro. Veamos un ejemplo:

Una alumna, Sofía, tiene notas en sus exámenes 8, 6, 7, 8 y 6, mientras que un alumno, Pedro, tiene notas 4, 10, 9, 5 y 7. Ambos tienen el mismo promedio - ¡calcúlalo! - , pero en el primer caso de de Sofía las notas son más consistentes, mientras que las notas de Pedro tienen más variabilidad. La **varianza** es, precisamente, una medida de esa variabilidad. Se calcula de la siguiente forma:

$$Var = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

donde X barra es el promedio de los datos. Si la fórmula te confunde, no te preocupes. Durante el encuentro explicaremos cómo interpretarla y cómo usarla. Pero, además, ¡ya viene programada en NumPy y en muchas otras librerías!

Otra medida muy utilizada es la **desviación estándar,** que es simplemente la raíz cuadrada de la varianza.

$$SD = \sqrt{Var}$$

Apliquemos estas fórmulas para las notas de Sofía y Pedro. En el primer caso, la varianza es 1, mientras que en el segundo es 5.2. Con la varianza podemos calcular la desviación estándar, 1 y 2.55 respectivamente. Podría parecer redundante tener dos medidas tan relacionadas, pero la ventaja de la desviación estándar es que tiene las mismas unidades que el promedio.

En general, **dado un conjunto de datos, la desviación estándar acompaña al promedio de los valores para darnos una idea de dispersión**, cuán lejos están de su promedio. En el ejemplo anterior, además del promedio de notas, nos gustaría saber si un/a estudiante en general se saca notas más o menos similares o si tiene altibajos pronunciados y, por ende, no es muy consistente su performance. En algunos contextos, podríamos informar las notas de Sofía como 7.0 ± 1.0, y las de Pedro como 7.0 ± 2.5. Ya ahí sabemos que ambos tienen el mismo promedio, pero Pedro fue menos consistente. Más adelante veremos más en detalle como la desviación estándar nos da aún más información,

Para pensar: ¿A qué se debe la variabilidad en los datos?

Veamos, entonces, cómo empezar a aplicar estas herramientas en Python y sobre un set de datos.

Una nueva librería

Pandas es una librería para manipular y analizar datos en Python. Está montada sobre NumPy, por lo cual muchas funcionalidades son similares. Utilizaremos Pandas para trabajar con datasets estructurados (y bueno, ¡bastante más!).

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

En el notebook que acompaña esta bitácora darás tus primeros pasos con esta librería, trabajando con el conjunto de Datos de Flores Iris. Si bien para los estándares de hoy en día este conjunto de datos no es particularmente desafiante, nos servirá para ejemplificar algunos usos de la librería de una manera clara. Además, este conjunto de datos es un clásico de Data Science, donde su uso como ejemplo ya se convirtió en legendario y sigue teniendo un gran valor histórico. Puedes leer más sobre este conjunto de datos en su entrada de Wikipedia y responder las siguiente preguntas:

- ¿Quién y cuándo construyó este dataset?
- ¿Qué es una flor iris?
- ¿Cuántas variables tiene?
- ¿Qué representa cada una?

Empieza a trabajar en el notebook

1. Instala Pandas

Instrucciones Pandas

2. Empieza a resolver el notebook

Una vez que esté instalado, trabaja en el siguiente notebook, ya sea en Colab o en sus computadoras. Para el encuentro, **lleva resuelta la Sección 1: "Estadística Descriptiva",** el resto del notebook lo trabajaremos en clase.

Notebook: DS_Bitácora_04_Estadística_y_Pandas

Puede descargar el dataset aquí. Si deseas trabajar en Colab, deberás subir el dataset. ¡Inténtalo, no es difícil!

¿Quieres saber más?

Profundiza
Te invitamos a conocer más sobre el tema de esta bitácora.

Comunidad
Sumérgete en la red donde los profesionales de DS interactúan.

Contenido extra del encuentro

DS-ONLINE-65 2/2/21
Tema: DS-ONLINE-65 2/2/21 12/8/21 Mar/Jue Hora de inicio de la reunión : 11 feb. 2021 06:54 Grabación de la reunión: https://acamica.zoom.us/rec/share/Z4vlE5Yhe60K458yzREmbZDYavZT9IvgjBcPIcoOP4lflGr9MniKWgrHyxfu_uEd.eJMQKssS13qi1J2r Código de acces