FEB
18

SPRINT 1 - ENCUENTRO 5

Probabilidad y Estadística - Pandas

Unirme

El encuentro comenzó hace
00:07:51

Lista de recursos

- Challenge
- Profundiza
- Notebook: DS_Bitácora_05_Probabili
- Potencia tu Talento - Comunidades

Conectando puntos

> ***"La estadística no sustituye el juicio." —Henry Clay Sr. Abogado y político estadounidense (1777-1852).***

En el encuentro y la bitácora anterior (Volver a las bases), comenzamos a ver algunos conceptos de Probabilidad y Estadística. En particular, vimos algunas herramientas de estadística descriptiva. Recordemos la situación que planteamos:

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?

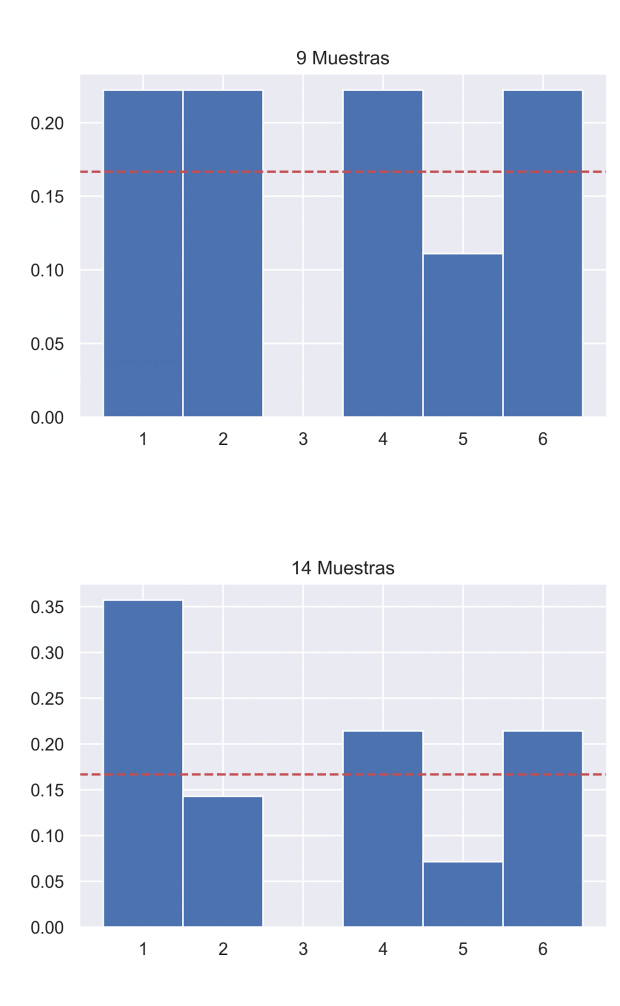


Figura 1: tomamos dos dados y anotamos los resultados de tirar cada uno. Luego, graficamos el porcentaje de veces que cae de cada lado. La línea punteada corresponde a lo esperado para un dado perfectamente balanceado. ¿Cuál será el que está cargado?

Siempre es importante que te tomes unos minutos para reflexionar sobre las situaciones que planteamos. Si ya estás listo/a, hoy veremos algunos conceptos fundamentales de Probabilidad, para terminar de complementar lo visto hasta acá.

Probabilidad

La Teoría de la Probabilidad es la disciplina que nos provee de herramientas para calcular y razonar acerca de probabilidades. Existen dos definiciones conocidas para probabilidad, no necesariamente excluyentes. Va a ser muy útil que las tengamos en cuenta ya que ambas sirven para pensar en nuestros problemas:

La definición más intuitiva de **Probabilidad** es la **Bayesiana:** es medida de la *confianza* o *certidumbre* de que un suceso ocurra. La mejor medida de la incertidumbre es la probabilidad. Sin embargo, ésta no suele ser la primera definición de probabilidad que estudiamos.

Veamos un ejemplo: Si un meteorólogo pronostica lluvia para mañana con una probabilidad de 0.99 (99% de probabilidad) y una meteoróloga pronostica lluvia para mañana pero con una probabilidad de 0.75 (75% de probabilidad), ¿quién tiene más confianza que lloverá al día siguiente?

Otra que también usamos mucho —sobre todo cuando tiramos monedas o dados— es la definición de **Probabilidad Frecuentista:** si hacemos un experimento muchas (!) veces, la probabilidad está asociada a la frecuencia con que ocurre cada posible valor de la variable aleatoria.

Veamos un ejemplo: Si tiramos muchas veces una moneda, ¿qué frecuencia (porcentaje) de caras y cecas esperamos? Si la moneda está balanceada esperamos que, en promedio, 50% de las veces caiga cara y 50% de las veces caiga ceca. Cuantas más veces tiremos la moneda, más cerca esperamos que estén de esos valores. Si pudiéramos tirar infinitas veces la moneda, la frecuencia con la que salen cara y ceca es la probabilidad de esos eventos.

Si bien hay interpretaciones que son más naturales en algunos contextos que otros, podríamos igualmente hacer una interpretación frecuentista del pronóstico para el día de mañana y una bayesiana para el lanzamiento de una moneda.

Al fin de cuentas, la probabilidad se materializa en un número, pero es importante que como data scientists tengamos en claro en qué está basada la interpretación que hacemos de la misma: por un lado, depende de la fuente (¿qué confianza tengo en los datos, en su análisis, etc.?) y, por otro, del contexto (¿cuántas veces podemos tirar un dado? ¿cuántas veces ocurrirá el día de mañana y su pronóstico meteorológico?). Por esto, es importante que tengamos ambas definiciones presentes y que sepamos cuándo y cómo utilizar cada una.

Independientemente de cuál definición te guste más, algunas propiedades de las probabilidades son válidas para ambas (notación: p(s) que refiere a la probabilidad 'p' de un resultado 's').

- *Las probabilidades son números mayores que cero y menores que uno. La notación es:* $0 \leq p(s) \leq 1$.

Ejemplo: en el caso de una moneda, $p(s = cara) = 0.5$, $p(s = ceca) = 0.5$.

- La suma de las probabilidades sobre todos los posibles resultados es uno.

Ejemplo: $p(s = cara) + p(s = ceca) = 0.5 + 0.5 = 1$

Para pensar: ¿cómo sería en el caso de las caras de un dado?

Ahora veremos algunas definiciones más, ¡que probablemente generen más confusión!

- Un **experimento** es un procedimiento que produce uno de un conjunto de posibles resultados. Tirar un dado, por ejemplo, es un experimento. También podemos considerar un experimento al trabajo que hace un/a médico/a cuando recibe en su consultorio a un nuevo paciente y anota en su historia clínica su edad, peso y altura. ¿Qué otras cosas pueden considerarse experimentos? ¿Cómo se relaciona esta definición con la acepción más coloquial de experimento?
- Un **espacio muestral** es el conjunto de posibles resultados de un experimento. Para un dado, el espacio muestral es $S = \{1,2,3,4,5,6\}$. Si tuviéramos dos dados, el espacio muestral sería todas las posibles combinaciones de los resultados de cada uno $S = \{(1,1), (1,2), (1,3), (1,4), \ldots, (2,1), (2,2), (2,3), \ldots, (6,3), (6,4), (6,5), (6,6)\}$. Es decir, un total de 36 combinaciones. ¿Cuál es el espacio muestral para: a) medir la temperatura de un objeto y b) tomar una persona al azar en la calle y preguntarle su edad?
- Una **variable aleatoria** es una función numérica sobre los resultados de un espacio de probabilidad. Como esta definición es confusa, informalmente podemos decir que es una variable aleatoria es una variable cuyos valores dependen del resultado de un suceso aleatorio. Por ejemplo, los resultados de tirar un dado conforman una variable aleatoria. Tirar dos dados y sumar sus resultados es otra variable aleatoria. **Nota:** en Probabilidad y Estadística, casi todo las variables que estudiamos son variables aleatorias, por lo que este concepto es central en estas áreas. Por ejemplo, la temperatura de un día o cuántos milímetros de lluvia cayeron son, desde este punto de vista, variables aleatorias. También puede serlo cuántos goles anota Lionel Messi en un partido de la Selección Argentina o en el Barcelona.

Para pensar: ¿que se llame variable aleatoria significa que está regida por el azar?

Distribuciones de Probabilidad

Tipos de datos numéricos

¿Recuerdas que en la bitácora anterior agrupamos las variables en cuantitativas (numéricas) y cualitativas (categóricas)?

El primer grupo corresponde a variables como, por ejemplo, edad, peso y altura de una persona, que pueden ser divididos en dos tipos: discretos o continuos. En tanto, el segundo grupo corresponde a variables como su color de pelo, género o profesión.

Si bien la diferencia puede ser sutil, hay que prestarle atención.

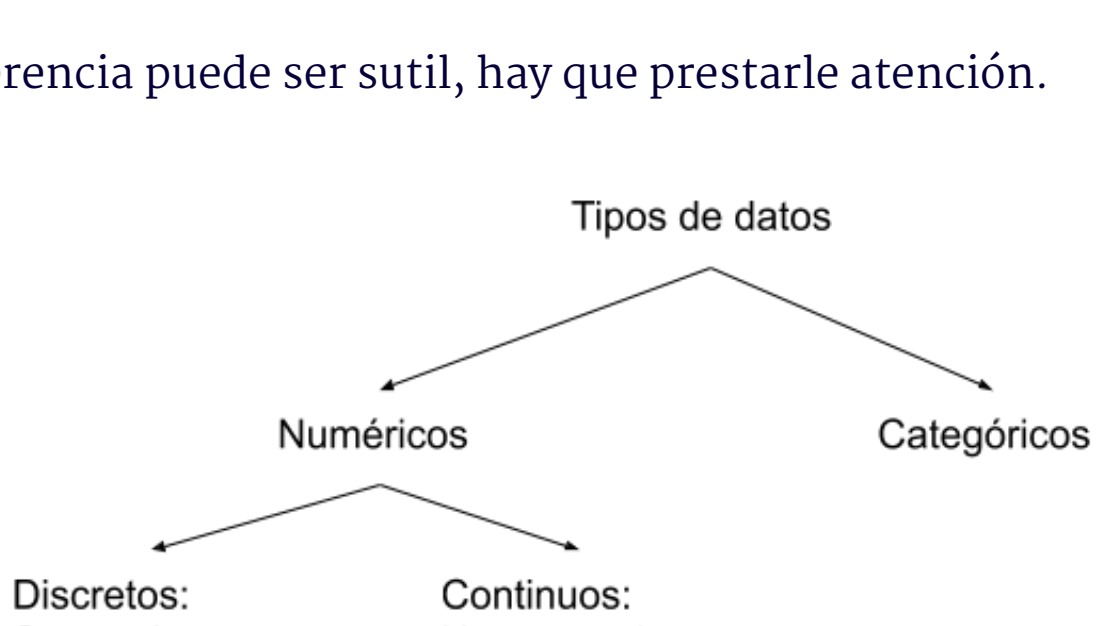


Figura 1: una posible taxonomía para los tipos de datos o variables estadísticos. No es la única taxonomía posible.

Las variables numéricas discretas son aquellas que se pueden contar, por ejemplo, la cantidad de hijos/as de una familia, cantidad de dormitorios en una casa, número de empleados de una empresa, resultado de tirar un dado, etc. Entre dos valores de una variable numérica discreta no existe otro valor. Las variables numéricas continuas son aquellas que, entre dos valores, siempre puede haber otro. Para pensarlo de otra forma, son aquellas que se miden. Por ejemplo, peso y altura de una persona, temperaturas, etc.

Otra cosa a la cual podemos prestarle atención es si una variable numérica (discreta o continua) está acotada, es decir, si existe un valor máximo o mínimo que puede tomar. Por ejemplo, las temperaturas están acotadas por debajo, ya que no puede haber temperaturas menores al cero absoluto; la cantidad de dormitorios de una casa no puede ser menor que cero; el resultado de tirar un dado de seis caras no puede ser menor que uno ni mayor que seis.

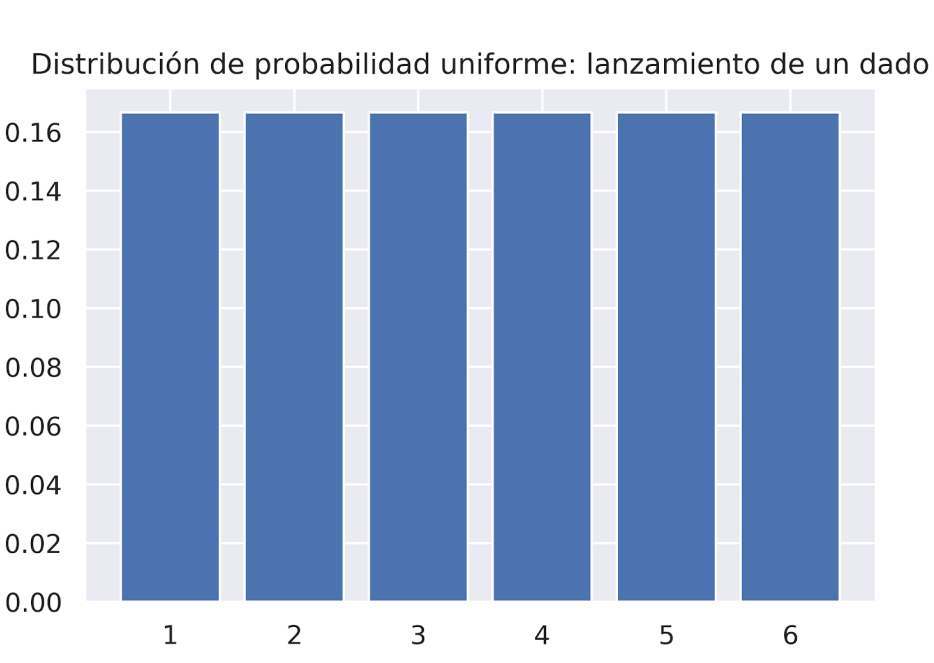
Para pensar: tiramos dos dados y anotamos la suma de sus resultados. ¿A qué tipo de variable corresponde? ¿Cuál es el valor máximo y mínimo que podemos obtener?

Distribución de probabilidad

Si estamos trabajando con una variable aleatoria, la mejor descripción que podemos obtener de ella es:

- su **densidad** de probabilidad, para el caso de una variable numérica continua (ej.: peso y altura de una persona).
- su **distribución** de probabilidad, para el caso de una variable numérica discreta. La **distribución de probabilidad** de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra.

Veamos un ejemplo: para el caso de un dado de seis caras, la distribución de probabilidad es uniforme, y asigna ⅙ de probabilidad a cada cara, como se ve en la figura debajo:



En el caso de una moneda, su distribución de probabilidad asigna ½ a cara y ½ a ceca:

A lo largo de las bitácoras, iremos viendo muchas distribuciones y densidades de probabilidad.

Relación entre Probabilidad y Estadística

Evidentemente, estas disciplinas están relacionadas, sino no aparecerían siempre juntas. Para ver esta relación, retomemos el ejemplo del dado que te propusimos analizar en la bitácora anterior. Todos tenemos en nuestra mente un modelo probabilístico de un dado de seis caras: asumimos que en un dado no cargado la probabilidad de que salga cada cara es ⅙. Pero para ver si esto se cumple, lo que tenemos que hacer es tirar muchas veces un dado y hacer estadística sobre sus resultados. Entonces, parecería ser que la Probabilidad se relaciona con lo que esperamos observar, mientras que la Estadística se relaciona con el análisis de lo que observamos.

Podríamos resumir ambas disciplinas como sigue:

- La **Probabilidad** se ocupa de cuantificar la certeza de eventos futuros, y crear modelos de lo que esperamos encontrar.
- La **Estadística** se ocupa de analizar eventos pasados. Se pregunta: lo que observamos, ¿tiene sentido con lo que esperábamos encontrar? Si no es así, ¿qué podemos aprender?

No te preocupes si la distinción no es del todo clara. Todos/as nos solemos confundir con esto.

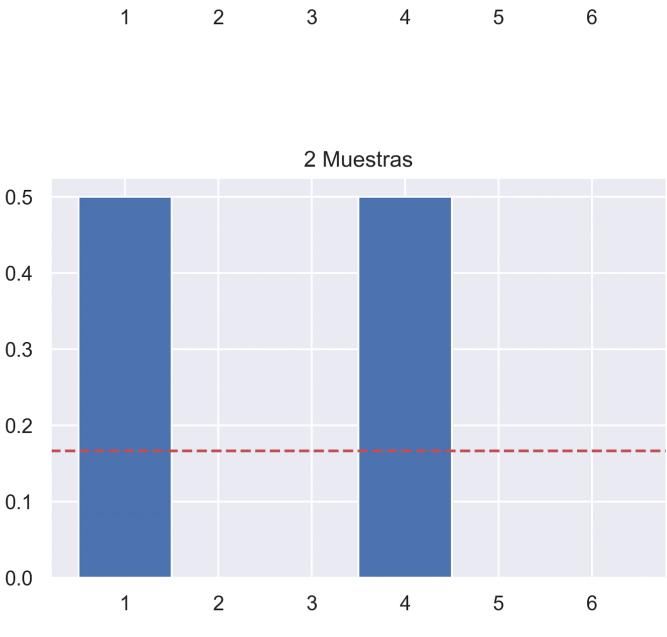
Para el caso del dado, la Probabilidad nos dice que, si el dado no está cargado, todas sus caras deberían ser igual de probables. Es decir, la distribución de probabilidad es uniforme y le asigna a cada cara una probabilidad de ⅙. En cambio, si el dado está cargado, esperamos que esa distribución sea distinta. Por ejemplo, si está cargado para favorecer al seis, que la probabilidad de ese valor sea mayor que un ⅙, a costa de disminuir la probabilidad del resto de los valores. La ciencia de la probabilidad nos ayuda a predecir lo que esperamos ver.

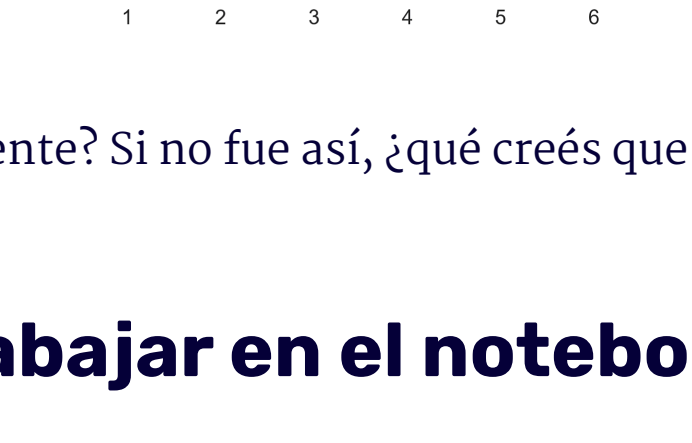
Una vez que ya tenemos una idea de lo que esperamos ver, debemos hacer los experimentos y hacer Estadística sobre sus resultados. Por ejemplo, tirando muchas veces los dados. En general, para cualquier experimento estadístico, podemos organizarnos de la siguiente manera:

- **Adquisición:** ¿qué datos necesito?¿En qué cantidad? Para el caso del dado, ¿qué valores anoto de los lanzamientos?¿Cuántos lanzamientos debo hacer?
- **Organización:** ¿cómo guardo y agrupo los datos obtenidos?
- **Análisis e Interpretación:** ¿en qué rango están estos valores? ¿Cómo son de precisos o de dispersos?¿Qué valores estadísticos puedo obtener de ellos que me ayuden a interpretarlos? Los datos, ¿se ajustan o no a lo esperado? Para el caso del dado, ¿qué dicen mis datos sobre si el dado está cargado o no?¿Hay alguna cara que haya salido más veces de lo esperado?
- **Presentación:** ¿cómo comunico mis conclusiones? ¿Qué números o gráficas utilizo para que se me entienda bien?

En el caso de los dos dados, los estadísticos que podemos obtener son relativamente sencillos: cantidad de veces que salió cada cara y su frecuencia. La ciencia de la estadística nos ayuda a analizar lo que obtuvimos.

Al comienzo de la bitácora te preguntamos cuál dado creías que estaba cargado. Acá está la respuesta:





¿Adivinaste correctamente? Si no fue así, ¿qué creés que falló?

Empieza a trabajar en el notebook

Trabaja en el siguiente notebook, ya sea en Colab o en tu computadora. Necesitarás instalar dos nuevas librerías, Scipy y Matplotlib, cuyo funcionamiento veremos más en detalle en la bitácora 06. ¡Se instalan como todas las librerías que instalamos hasta el momento! Para el encuentro, lleva hecha la sección 1, el resto del notebook lo trabajaremos en clase.

Notebook: DS_Bitácora_05_Probabilidad

¡Prepárate para el próximo encuentro!

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

Challenge
Te proponemos el siguiente desafío, ¿te animas?

Potencia tu Talento - Comunidades
Tips para construir tu perfil profesional.